World Health Organization

REGIONAL OFFICE FOR Europe

Tools for making good data visualizations: the art of charting

World Health Organization

REGIONAL OFFICE FOR Europe

Tools for making good data visualizations:

the art of charting

Abstract
Data visualization is a collection of methods that use visual representations to explore, make sense of and communicate quantitative data. It allows trends and patterns in quantitative data to be seen more easily. The ultimate purpose of data visualization is to facilitate better decisions and actions. With the rise in the amount of data available, it is important to be able to interpret increasingly large batches of information, and data visualization is a strong tool for the job. Use of good data visualizations is essential to make impactful health reports. This tool provides practical guidance on how to make good data visualizations to support convincing policy messages.

This guidance document is part of the WHO Regional Office for Europe's work to support Member States in strengthening their health information systems. Helping countries to produce solid health intelligence and institutionalized mechanisms for evidence-informed policy-making has traditionally been an important focus of WHO's work and continues to be so under the European Programme of Work 2020–2025.

Keywords:
DATA VISUALIZATION, DATA DISPLAY, COMMUNICATION, DECISION MAKING

Document number: WHO/EURO:2021-1998-41753-57181

© World Health Organization 2021
Some rights reserved. This work is available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO licence (CC BY-NC-SA 3.0 IGO; https://creativecommons.org/licenses/by-nc-sa/3.0/igo).

Under the terms of this licence, you may copy, redistribute and adapt the work for non-commercial purposes, provided the work is appropriately cited, as indicated below. In any use of this work, there should be no suggestion that WHO endorses any specific organization, products or services. The use of the WHO logo is not permitted. If you adapt the work, then you must license your work under the same or equivalent Creative Commons licence. If you create a translation of this work, you should add the following disclaimer along with the suggested citation: "This translation was not created by the World Health Organization (WHO). WHO is not responsible for the content or accuracy of this translation. The original English edition shall be the binding and authentic edition: Tools for making good data visualizations: the art of charting. Copenhagen: WHO Regional Office for Europe; 2021".

Any mediation relating to disputes arising under the licence shall be conducted in accordance with the mediation rules of the World Intellectual Property Organization (http://www.wipo.int/amc/en/mediation/rules/).

Suggested citation. Tools for making good data visualizations: the art of charting. Copenhagen: WHO Regional Office for Europe; 2021. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Sales, rights and licensing. To purchase WHO publications, see http://apps.who.int/bookorders. To submit requests for commercial use and queries on rights and licensing, see http://www.who.int/about/licensing.

Third-party materials. If you wish to reuse material from this work that is attributed to a third party, such as tables, figures or images, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.

General disclaimers. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of WHO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers' products does not imply that they are endorsed or recommended by WHO in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by WHO to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall WHO be liable for damages arising from its use.

Contents

Acknowledgements

This document was developed by the Data, Metrics and Analytics Unit in the Division of Country Health Policies and Systems of the WHO Regional Office for Europe. The main author is Laurens Zwakhals. Marieke Verschuuren and David Novillo Ortiz provided direction during the production of the report and technical advice during concept drafting, writing and review. Special thanks to Natasha Azzopardi-Muscat for her strategic guidance.

For further information please contact the Data, Metrics and Analytics Unit (euhiudata@who.int).

Aim of this guidance

This guidance document is part of the WHO Regional Office for Europe's work on supporting Member States in strengthening their health information systems. Helping countries to produce solid health intelligence and institutionalized mechanisms for evidence-informed policy-making has traditionally been an important focus of WHO's work, and continues to be so under the European Programme of Work 2020–2025.[1] Use of good data visualizations is essential to make impactful health reports. This tool provides practical guidance on how to make good data visualizations to support convincing policy messages.

1 European Programme of Work. In: WHO Regional Office for Europe [website]. Copenhagen: WHO Regional Office for Europe; 2020 (https://www.euro.who.int/en/health-topics/health-policy/european-programme-of-work/european-programme-of-work, accessed 19 November 2020).

Note on the figures

All figures in this guidance document were created by the author except Fig. 15, which has a Creative Commons (CC-BY 4.0) licence and is free to share, copy and redistribute in any medium or format. Data sources are cited below each figure.

Introduction: data visualization and why it is important

Data visualization is a collection of methods that use visual representations to explore, make sense of and communicate quantitative data *(1)*. It allows trends and patterns in quantitative data to be seen more easily. The ultimate purpose of data visualization is to facilitate better decisions and actions.

With the rise in the amount of data available, it is important to be able to interpret increasingly large batches of information, and data visualization is a strong tool for the job. This is not only important for data scientists and data analysts: it is necessary to understand data visualization in any career. Anyone who works in finance, marketing, design, health monitoring or any other sector needs to visualize data. This showcases the importance of data visualization *(2)*.

Data visualization can be used in various phases of data analysis and communication. First, it is very helpful in the phase of **data exploration**. With the help of statistical analysis tools and spreadsheet programs, data can easily be analysed through visualization. Relationships, distributions and comparisons can be discovered. Business intelligence tools are also extremely helpful for gaining insight into the data. In the phase of data exploration, the choice of chart type, side information and layout are not the most important issues: as long as the visuals give the analyst more insight, that is fine.

Once a good understanding of the trends and patterns in the data has been developed, the work moves on to think of the story to be told. In this phase, **data presentation** comes into play. Data presentation (sometimes called data explanation) aims to report the end results. This can be in a book, a report, a presentation or on the web. When transmitting insights to a broad audience, clear and understandable communication becomes important. In this phase the chart type, side information and layout of the visuals become important. This guidance document is about the use of data visualizations in this phase of the process: moving from data via information to insight for a broader audience. The visuals need to speak for themselves.

Reader’s guide

- This guidance document starts with the basics. Section 1 discusses why data visualizations are needed. An example shows that data visualization adds insight to non-visual analyses. The section further explains how visuals add insight about relationships, distributions and comparisons.
- Section 2 discusses some characteristics of datasets. Depending on these, different charts may be appropriate.
- Section 3 gives an overview of the most used charts. The options are endless, but the 13 types discussed in this section cover those used in 90% of cases.
- Section 4 examines the side information. A chart consists of both graphics and textual elements: the textual elements give the graphics context. This side information is important for understanding what the chart is about.
- Section 5 investigates obvious and less obvious pitfalls in charting data.
- Finally, section 6 discusses some layout principles. Especially when more than one chart is used in a document, a unifying layout is important for the audience. Once the reader is used to a specific layout, they will find interpretation of the next chart easier.

1. Why data visualization is needed

Data visualization is needed because a visual summary of information makes it easier to identify patterns and trends than looking through thousands of rows on a spreadsheet. It is the way the human brain works. Since the purpose of data analysis is to provide insight, data are much more valuable when visualized. Even if a data analyst can identify the patterns and pull insight from data without visualization, it is more difficult to communicate the data findings and their meaning without charts.

1.1 The power of visualization: an example

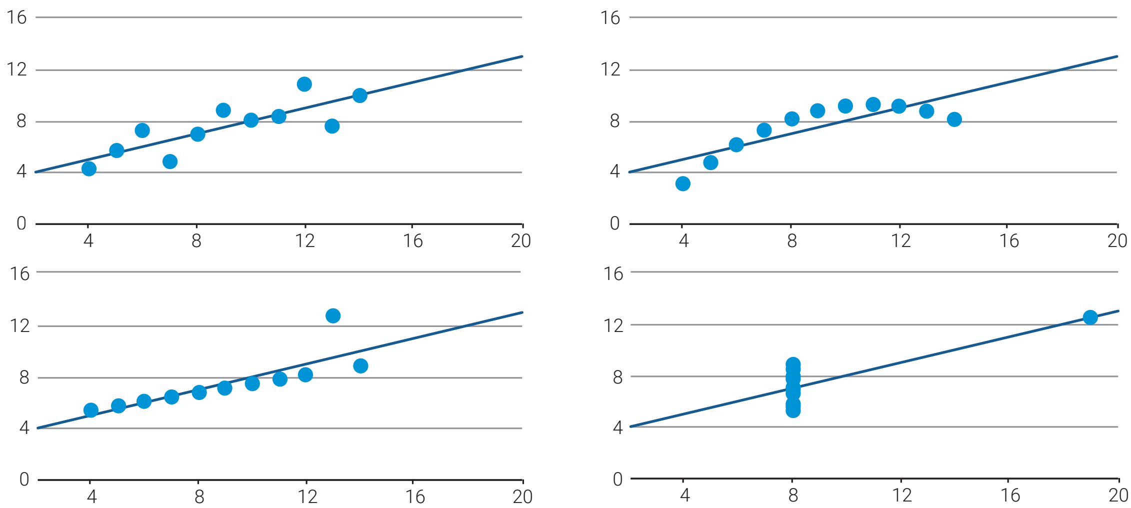
The power of visualization can be illustrated by a simple example known as **Anscombe's quartet** *(3)*. It comprises four datasets that have nearly identical simple descriptive statistics, yet have different distributions and appear vastly different when charted. Each dataset consists of 11 (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate the importance of charting data before analysing it, to counter the impression among statisticians that "numerical calculations are exact, but charts are rough".

Table 1 shows the datasets Anscombe constructed. The x values are the same for the first three datasets. At the end of the table several descriptive statistics are shown. Essential statistics are the same, so it is possible to conclude that the four datasets are similar, but when they are charted (Fig. 1), the differences become clear at a glance. Visualizing data can provide insight that tabular data and descriptive statistics do not always give.

Table 1. Anscombe's quartet

	Dataset I		Dataset II		Dataset III		Dataset IV	
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
Summary statistics								
Number	11	11	11	11	11	11	11	11
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Standard deviation	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
Correlation coefficient	0.82		0.82		0.82		0.82	

Fig. 1. Anscombe's quartet visualized



Data source: Anscombe (3).

Without a visual representation of the insights, it can be hard for the audience to grasp the true meaning of the findings. For example, showing figures of population life expectancy for the past 20 years will not tell an audience why they should care about the data, but showing a graph of the growth or decline of this life expectancy – preferably compared to other countries' data – will be sure to get their attention.

1.2 Uses of data visualization

Data visualization has many uses, and each type can be used in different ways. Section 3 will cover the various types in detail; this section sets out some of the most common ways data visualization is used.

1.2.1 Changes over time

This is perhaps the most basic and common use of data visualization, and may therefore be the most valuable. It is the most common because most data have an element of time involved. Therefore, the first step in a lot of data analyses is to find out the data trends over time. **Line charts** are best used for this type of data.

If many trendlines need to be visualized, line charts tend to become cluttered and form **spaghetti** plots. Colour can be added as an extra graphic variable. Data visualization can be as simple as colour coding; this can help when visualizing big datasets with many data lines. The colour codes simply add an extra variable on top of the location on the X and the Y axes. A **heatmap** is the manifestation of this solution.

1.2.2 Determining frequency

Showing frequency is also a much-used function of data visualization. It applies to data that are divided into classes; these can be either ordinal or nominal. **Column charts**, **bar charts** and **histograms** are used for this type of data visualization.

1.2.3 Determining relationships (correlations)

Identifying correlations is an extremely valuable use of data visualization. It is difficult to determine the relationship between two variables without a visualization, yet it is important to be aware of relationships in data. **Scatterplots** and **bubble charts** are extremely helpful for this.

The visualization of Anscombe's datasets (Fig. 1) is an example of a scatter plot. This is a great example of the value of data visualization in data analysis. Even descriptive statistics are insufficient to get insight into correlations and outliers.

1.2.4 Examining spatial patterns

If a variable has a spatial component, a good way to visualize the distribution over the area (such as a country, region or continent) is a **map**. This may be the right option for visualizing the variation of mortality between regions in a country, for example.

Another example in which a map is a good option is the need to visualize pandemic risk through the degree of connectedness. Epidemiologists use this type of visualization to analyse the way an epidemic is spreading over the country, creating **flow maps** for this purpose.

1.2.5 Analysing values and confidence

Determining complex metrics – such as values including confidence intervals – requires many different variables to be factored in, making the data almost impossible to see accurately with a plain spreadsheet. To visualize values with confidence levels an **area range chart** is a strong tool.

1.2.6 Combinations

Sometimes a combination of ways to visualize data is needed. If, for instance, the change of frequency over time needs to be visualized, two bar charts can be presented side by side, but a **slope chart** would be the better option. Comparison of frequencies with another population or target can be done with a **bullet chart**.

2. Characteristics of the data

Every dataset has various characteristics. This section discusses two characteristics that are important to consider when choosing the right chart: whether the dataset has individual cases or grouped cases and the measurement scale.

2.1 Individual versus grouped data

The first characteristic of a dataset is whether it contains individual or frequency data. For example, individual data can be visualized in a **scatter plot** or can be summarized as frequencies in order to be visualized as a **histogram**.

The main characteristic of grouped data is that datapoints are classified into groups. Examples include:

- number of new COVID-19 cases per day
- number of people grouped in body mass index classes
- number of people grouped by cause of mortality.

Grouped data are best visualized with **line charts**, **column charts** or **bar charts**.

2.2 Level of measurement

The second characteristic of a dataset is its level of measurement. In the 1940s, Stanley Smith Stevens introduced four levels of measurement: **nominal**, **ordinal**, **interval** and **ratio** *(4)*. These are still widely used today as a way to describe the characteristics of a variable. Knowing the level of measurement for a variable is not only an important aspect of choosing the right statistical analysis but also an important aspect of choosing the right chart to visualize the data.

2.2.1 Nominal

A nominal scale describes a variable with categories that do not have a natural order or ranking: a **qualitative** scale. Components consist of associated and differentiated groups, yet have no inherent order, like a list of the 50 states in America. Nominal variables can be coded with numbers, but the order is arbitrary, and any calculations – such as computing a mean, median or standard deviation – would be meaningless.

Examples of nominal variables include:

- blood type
- zip code
- gender
- race
- eye colour
- political party
- religion
- the countries in the world.

In general a **bar chart** is the best visualization to use for this scale of measurement. As qualitative components have no inherent order, they can be re-ordered arbitrarily to reveal patterns in the data.

2.2.2 Ordinal

An ordinal scale is one where the **ranking order** matters but not the difference between values. Components consist of elements with a natural sequence, like cold-warm-hot or white-grey-black.

Examples of ordinal variables include:

- socioeconomic status (low income, middle income, high income)
- education level (primary, secondary, bachelor’s degree, master’s degree, doctorate)
- income level (less than 50 000, 50 000–100 000, over 100 000)
- satisfaction rating (dislike intensely, dislike, neutral, like, like intensely).

Note that the differences between adjacent categories do not necessarily have the same meaning. For example, the difference between the two income levels “less than 50 000” and “50 000–100 000” does not have the same meaning as the difference between the two income levels “50 000–100 000” and “over 100 000”.

In general, a **column chart** is the best visualization to use for this scale of measurement.

2.2.3 Interval

An interval scale is one where there is order and where the difference between two values is meaningful. Components consist of elements with constant numerical ratios between one another, like a sequence of minutes.

Examples of interval variables include:

- temperature (Fahrenheit)
- temperature (Celsius)
- acidity (pH value).

2.2.4 Ratio

A ratio variable has all the properties of an interval variable, but also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable.

Examples of ratio variables include:

- duration
- weight
- length
- dose amount
- temperature in Kelvin (0.0 Kelvin really does mean “no heat”)
- survival time
- pulse.

When working with ratio variables, but not interval variables, the ratio of two measurements has a meaningful interpretation. For example, because weight is a ratio variable, a weight of 4 g is twice as heavy as a weight of 2 g. A temperature of 10 degrees Celsius should not be considered twice as hot as 5 degrees Celsius, however. If it were, a conflict would be created because 10 degrees Celsius is 50 degrees Fahrenheit and 5 degrees Celsius is 41 degrees Fahrenheit. Clearly, 50 degrees is not twice 41 degrees. In another example, a pH of 3 is not twice as acidic as a pH of 6, because pH is not a ratio variable.

For both interval and ratio variables, **line charts** are the best visualization to use.

3. Types of chart

The previous sections set out how data visualizations can be used and some of the characteristics of the underlying data. Section 3 moves on to applying the different types of data visualization to their uses. Numerous tools are available to help create data visualizations. Some are more manual and some are automated, but either should assist with creation of any of the types of visualization discussed in this section.

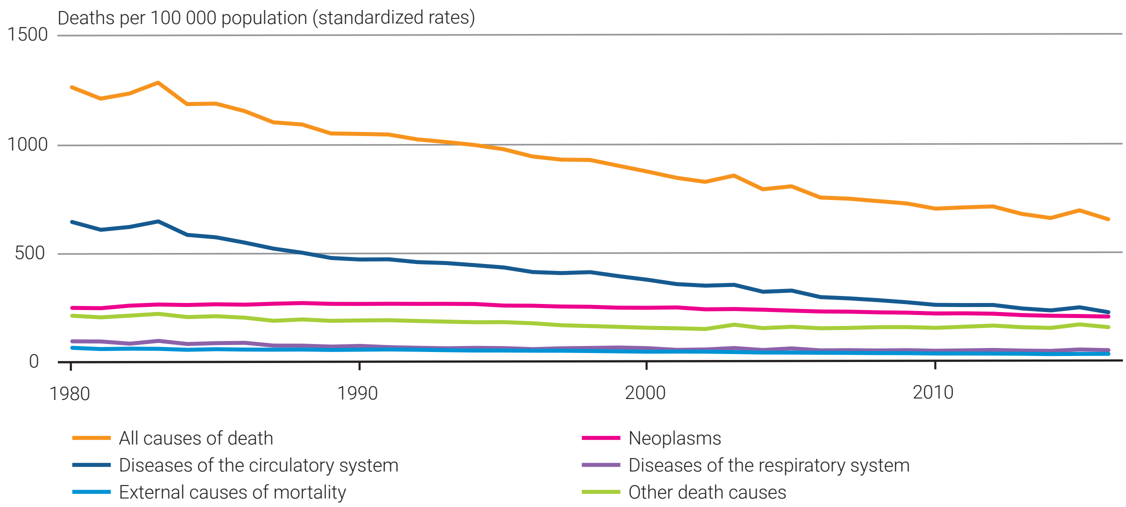
This section examines 13 different chart types. Many more are available (such as gauges, tree maps or frame diagrams, column ranges, packed bubble charts, box plots and Sankey diagrams), but the chart types discussed below should meet over 90% of data visualization needs.

3.1 Line chart

A line chart is used for interval or ratio scales on the X axis, while the Y axis is quantity. A timescale on the X axis is common, but other continuous scales are also possible. A line chart is very helpful for illustrating changes in mortality by cause of death, as shown below in Fig. 2.

Fig. 2. Example of a line chart

Causes of death in Italy, 1980-2016

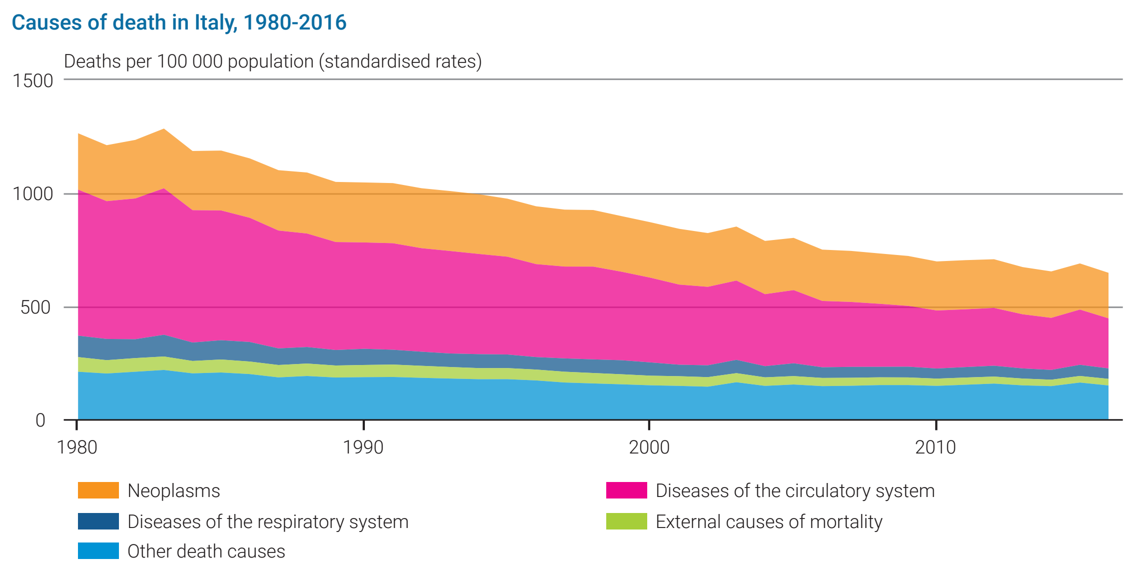


Data source: OECD.stat [online database]. Paris: Organisation for Economic Co-operation and Development; 2020 (https://stats.oecd.org/, accessed 18 December 2020).

3.2 Area chart

An area chart is an adaptation of a line chart where the area under the line is filled in to emphasize its significance. If more than one variable is shown, an area chart needs to be used as a **stacked line chart**, as shown in Fig. 3; the values for every category shown in the chart add up to the total. It is not necessary to show the "All causes of death" line in this chart type, since the individual stacks add up to the totals.

Fig. 3. Example of an area chart



Data source: OECD.stat [online database]. Paris: Organisation for Economic Co-operation and Development; 2020 (https://stats.oecd.org/, accessed 18 December 2020).

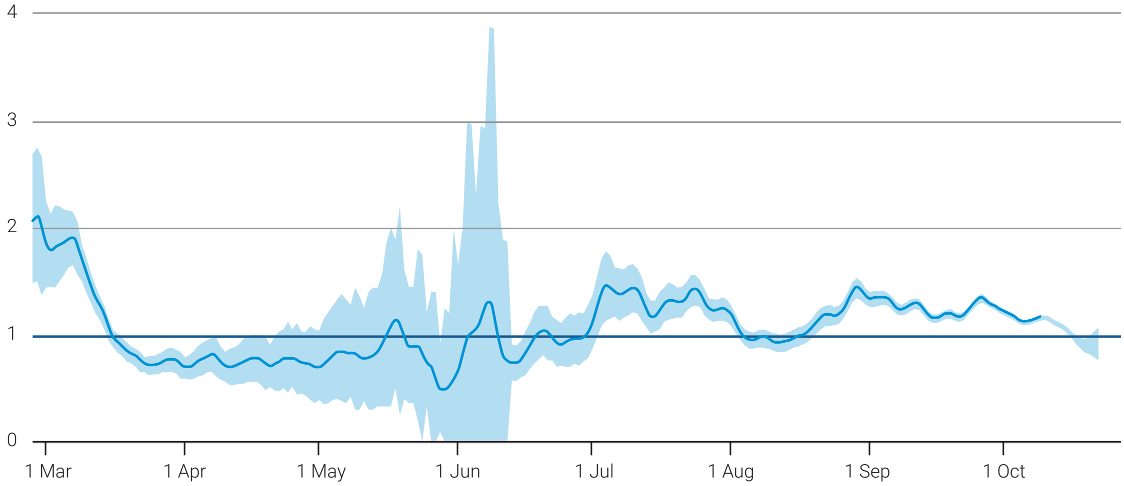
3.3 Area range chart

An area range chart is an area chart where the area range is defined with both low and high values. Only the area between the low and high values is filled in. The area range chart type is primarily used to show the estimated range of a specific indicator.

An area range chart is often combined with a **line chart**, showing the average line between the low and high values (see Fig. 4). This example shows the estimated COVID-19 reproduction rate in the Netherlands along with the significance interval visualized as an area range.

Fig. 4. Example of a range area chart combined with a line chart

COVID-19 reproduction rate 2020, The Netherlands



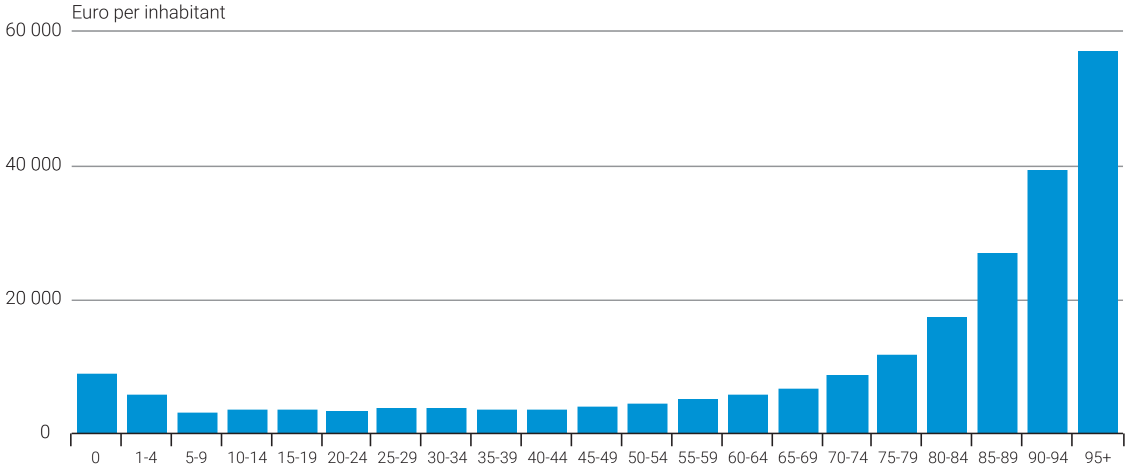
Data source: Covid-19 reproductiegetal [online database] [in Dutch]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2020 (https://data.rivm.nl/geonetwork/srv/dut/catalog.search#/metadata/ed0699d1-c9d5-4436-8517-27eb993eab6e, accessed 18 December 2020).

3.4 Column chart

A column chart is the best visualization to use to show a distribution of values in ordinal variables. Fig. 5 is an example of the distribution of health care expenditure by age category.

Fig. 5. Example of a column chart

Healthcare expenditure by age in the Netherlands, 2015



Data source: Kosten van ziekten 2015 [online database] [in Dutch]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2019 (https://statline.rivm.nl/#/RIVM/nl/dataset/50040NED/table?dl=481DC, accessed 18 December 2020).

3.5 Bar chart

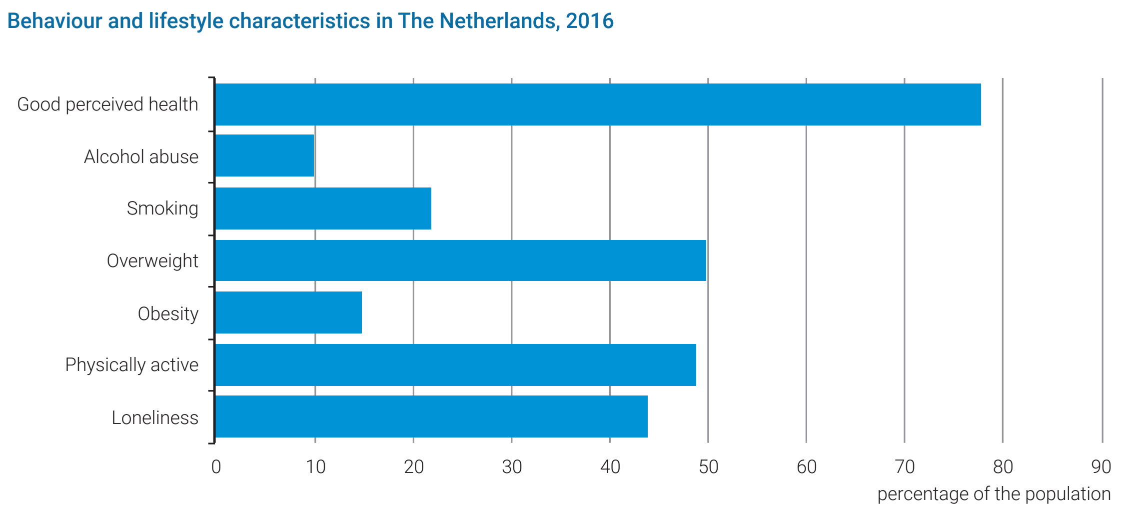
A bar chart is best for showing a distribution of values in nominal variables. Note that no labels can be skipped on the Y axis in a bar chart. This is typical for a nominal variable. For ordinal variables in a column chart like Fig. 5, labels on the X axis can be skipped, although this is not ideal.

Since the label names in nominal variables can be long, a bar chart is of greater use than a column chart. By right-aligning the labels on the Y axis, the data can be decoded more easily.

If nominal data labels are charted in a column chart, many tools automatically rotate text on the X axis when the labels become too long to fit horizontally. As well as looking messy, diagonally rotated text is harder to read, so a bar chart is better for nominal data.

Technically, in a bar chart the X axis and Y axis are switched so that the Y axis is horizontally oriented while the X axis is vertically oriented.

Fig. 6. Example of a bar chart

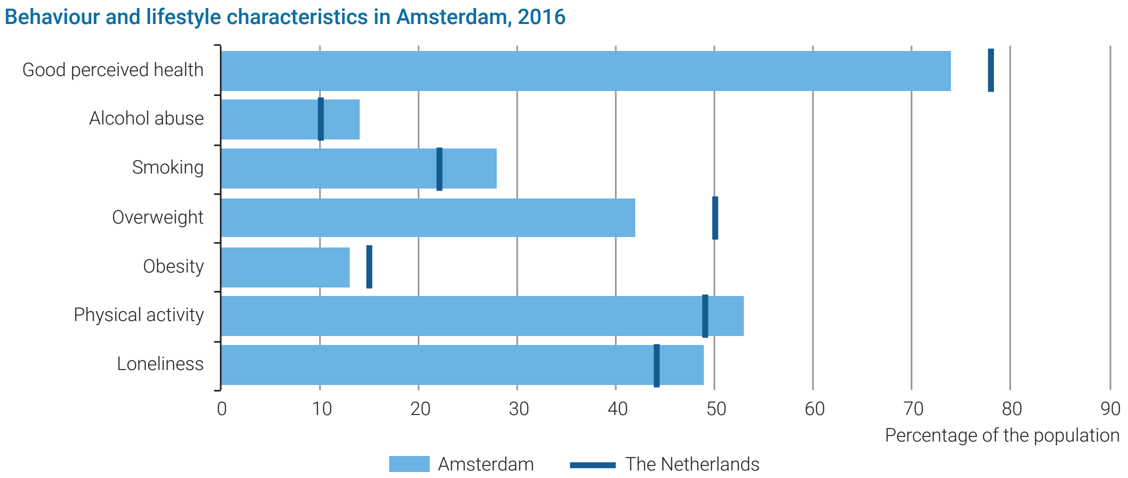


Data source: Gezondheid per wijk en buurt 2016 [online database] [in Dutch]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2019 (https://statline.rivm.nl/#/RIVM/nl/dataset/50052NED/table?dl=481DE accessed 18 December 2020).

3.6 Bullet chart

A bullet chart is a bar chart with an extra indicator on top of every bar. These lines are called **bullets**. They may show, for example, the values of another population group or target values for comparison.

Fig. 7. Example of a bullet chart



Data source: Gezondheid per wijk en buurt 2016 [online database] [in Dutch]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2019 (https://statline.rivm.nl/#/RIVM/nl/dataset/50052NED/table?dl=481DE accessed 18 December 2020).

3.7 Histogram

A histogram looks like a bar chart but measures frequency rather than a trend over an ordinal scale. The X axis of a histogram lists the **bins** or intervals of the variable; the Y axis shows frequency, so each bar represents the frequency of that bin.

A histogram is an approximate representation of the distribution of numerical data. It was first introduced by Karl Pearson *(5)*. To construct a histogram, the first step is to bin the range of values – that is, divide the entire range of values into a series of intervals – and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. They must be adjacent, and are often (but are not required to be) of equal size *(6)*.

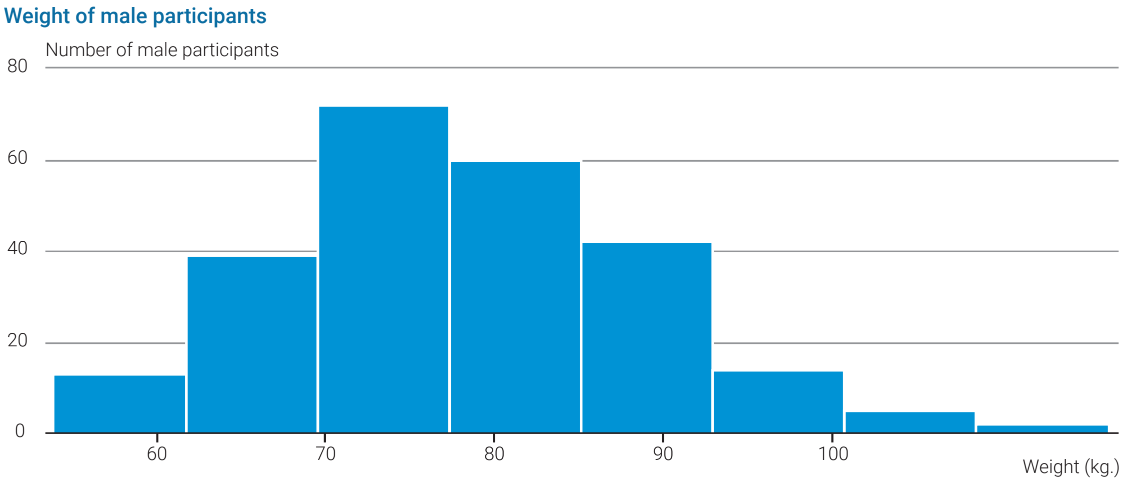
Because a histogram shows frequency, and therefore visualizes the **distribution** of a variable, the columns are usually drawn back to back with no space between them.

The bins can be constructed in numerous ways. There is no "best" number, and different bin sizes can reveal different features of the data. Using wider bins where the density of the underlying data points is low reduces noise due to sampling randomness; using narrower bins where the density is high (so that the signal drowns the noise) gives greater precision to the density estimation. Thus, varying the bin width within a histogram can be beneficial. Nonetheless, equal-width bins are widely used.

Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution. Depending on the actual data distribution and the goals of the analysis, different bin widths may be appropriate, so experimentation is usually needed to determine an appropriate width. There are, however, various useful guidelines and rules of thumb.

The most common way to select the number of bins is the **square root choice**. This takes the square root of the number of data points in the sample and rounds to the next integer. This method is used by Excel histograms and many others. Wikipedia is a good start for getting insight in selecting other methods *(6)*.

Fig. 8. Example of a histogram



Data source: artificially generated dataset.

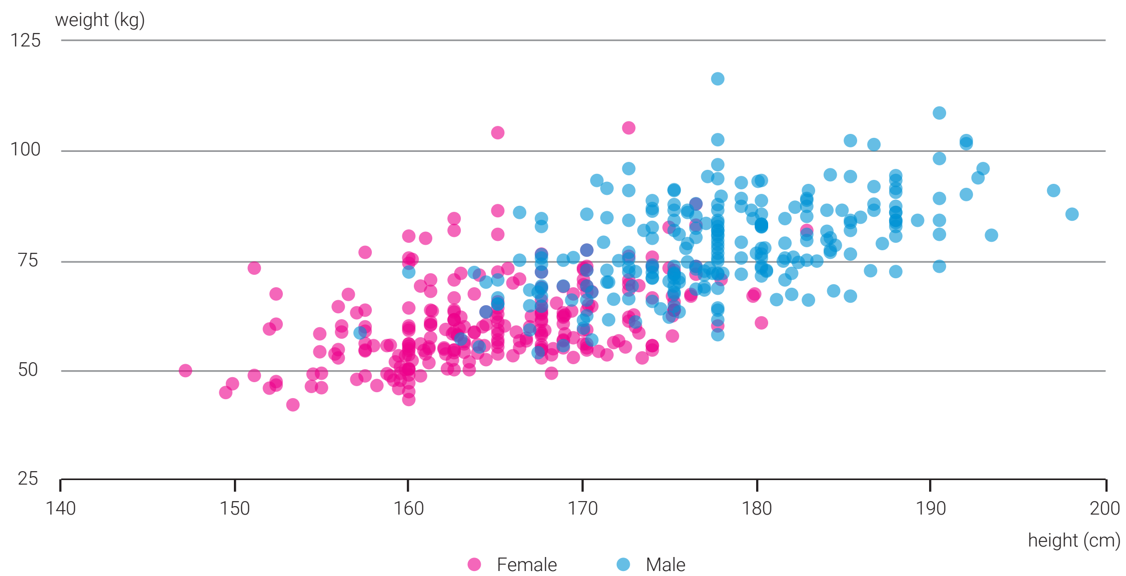
3.8 Scatter plot

Scatter plots are used to find correlations. Each point on a scatter plot means "when x = this, then y equals this". That way, if the points trend a certain way (upward to the left, downward to the right, and so on) there is a relationship between them. If the plot is truly scattered with no trend, then the variables do not affect each other at all.

The scatter plot of Fig. 9 shows the relationship between weight and height of 507 individuals by gender, from a study exploring relationships in body dimensions *(7)*.

Fig. 9. **Example of a scatter plot**

Height versus weight of 507 individuals by gender



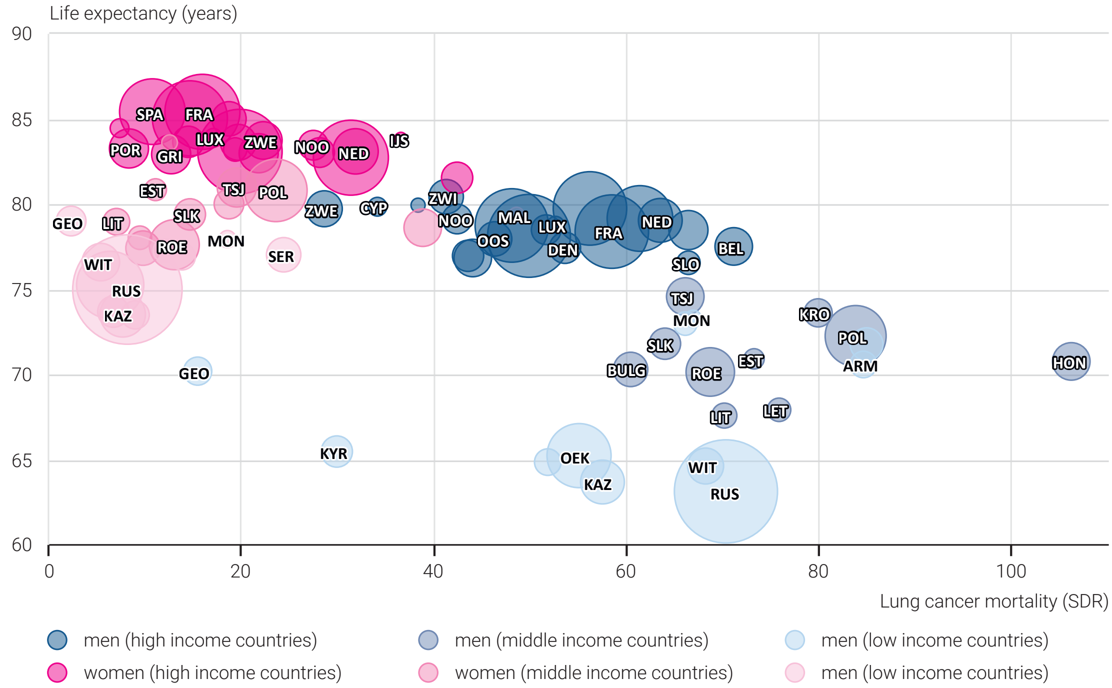
Data source: Heinz et al. *(7)*.

3.9 Bubble chart

A bubble chart is an adaptation of a scatter plot, where each point is illustrated as a bubble whose area has meaning in addition to its placement on the axes. A problem associated with bubble charts is the limitation on sizes of bubbles due to the limited space within the axes: not all data will fit effectively in this type of visualization.

Fig. 10. Example of a bubble chart

Life expectancy by lung cancer mortality in countries of the WHO European Region by GDP country clusters, 2010



Data source: European Health Information Gateway: Health for All explorer [online database]. Copenhagen: WHO Regional Office for Europe; 2020 (https://gateway.euro.who.int/en/hfa-explorer/, accessed 18 December 2020).

3.10 Pie chart

A pie chart is an option for illustrating percentages, because it shows each element as part of a whole. Its primary strength is the fact that the "part-to-whole relationship" message is built right into it in an obvious way *(8)*.

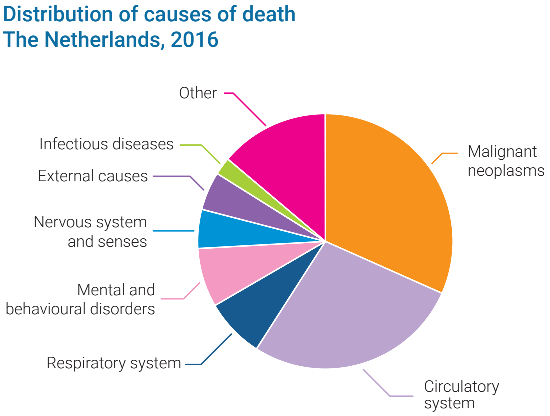
Despite the obvious nature of a pie chart's message, however, bar charts provide a much better means to compare the magnitudes of each part. Pie charts only make it easy to judge the magnitude of a slice when it is close to 0%, 25%, 50%, 75% or 100%. Any percentages other than these are difficult to discern in a pie chart, but can be accurately discerned in a bar chart, thanks to the quantitative scale *(8)*.

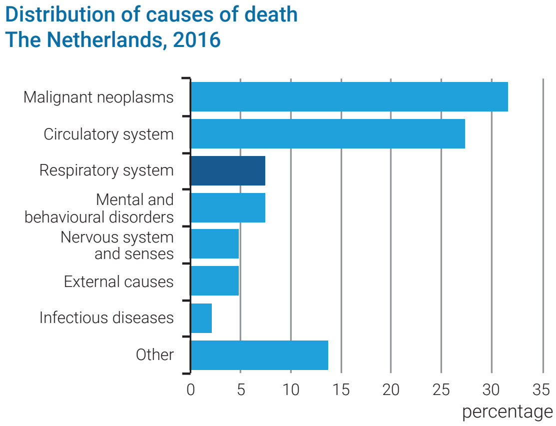
A bar chart is the preferred alternative to a pie chart. Fig. 11 shows the same data set visualized in both ways. In this example, it is difficult to read the percentage of people that die from failure of the respiratory system in the pie chart. The bar chart makes it much easier, showing at a glance that around 7.5% of total mortality is caused by failure of the respiratory system.

A pie chart needs different colours for different slices, meaning that numerous colours may be needed to visualize the data. In a bar chart the different elements do not have to have different colours, although attention can be drawn to one element by highlighting that bar with a different colour, as in the bar chart of Fig. 11.

For further evidence that a pie chart may not be the best visualization, see Stephen Few's essay *Save the pies for dessert (8)*.

Fig. 11. Example of a pie chart and a bar chart as a suggested alternative





Data source: Levensverwachting. In: VTV-2018 [website] [in Dutch]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2018 (https://www.vtv2018.nl/Levensverwachting#bv1_3_1, accessed 18 December 2020).

3.11 Slope chart

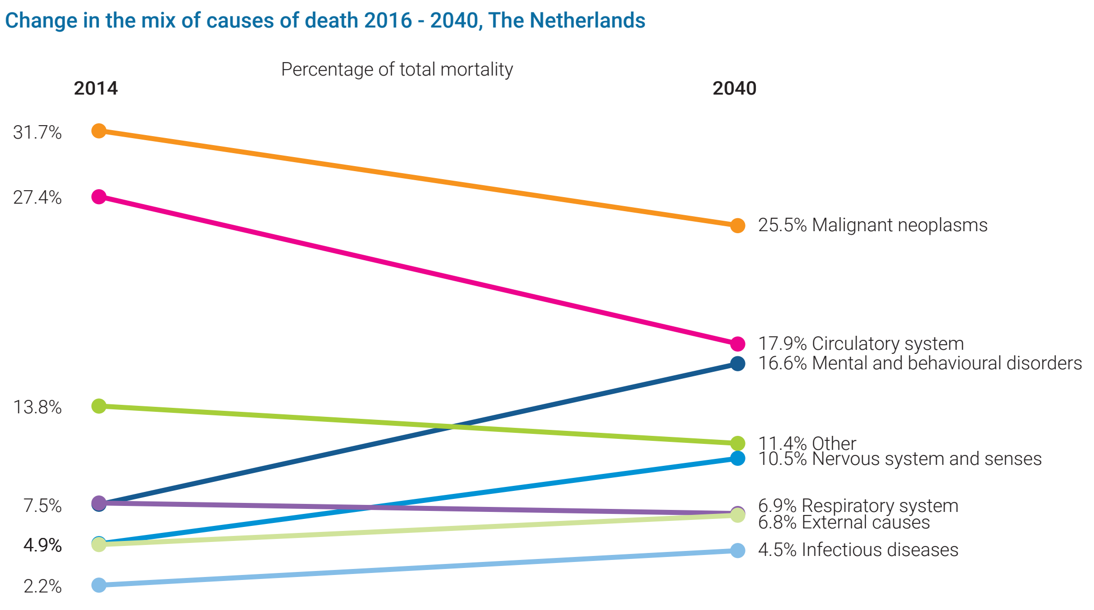
Slope charts are the line chart's fraternal twin. Line charts display three or more points in time, while slope charts display exactly two points in time.

Slope charts are really powerful for visualizing changes over time in a set of entities (such as countries, causes of death or energy sources). They show the change not only within one entity but also compared to the other entities in the set.

Slope charts are defined by Edward Tufte in his 1983 book *The visual display of quantitative information (9)*. This type of chart is useful for showing:

- the hierarchy of a set of entities to be compared at two moments in time (in the example of Fig. 12: the order of causes of death in both 2014 and 2040 (projected));
- the specific percentage associated with each entity in each of those years (the percentage next to the causes of death);
- changes over time for each entity (the slope of each cause of death);
- how each entity's rate changes compared to the other entities' rates of change (the slopes compared with one another);
- any notable deviations in the general trend (notice the change in the Mental and behavioural disorders line, compared to the other causes of death): aberrant slopes.

Fig. 12. Example of a slope chart



Data source: Levensverwachting. In: VTV-2018 [website] [in Dutch]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2018 (https://www.vtv2018.nl/Levensverwachting#bv1_3_1, accessed 18 December 2020).

3.12 Heatmap

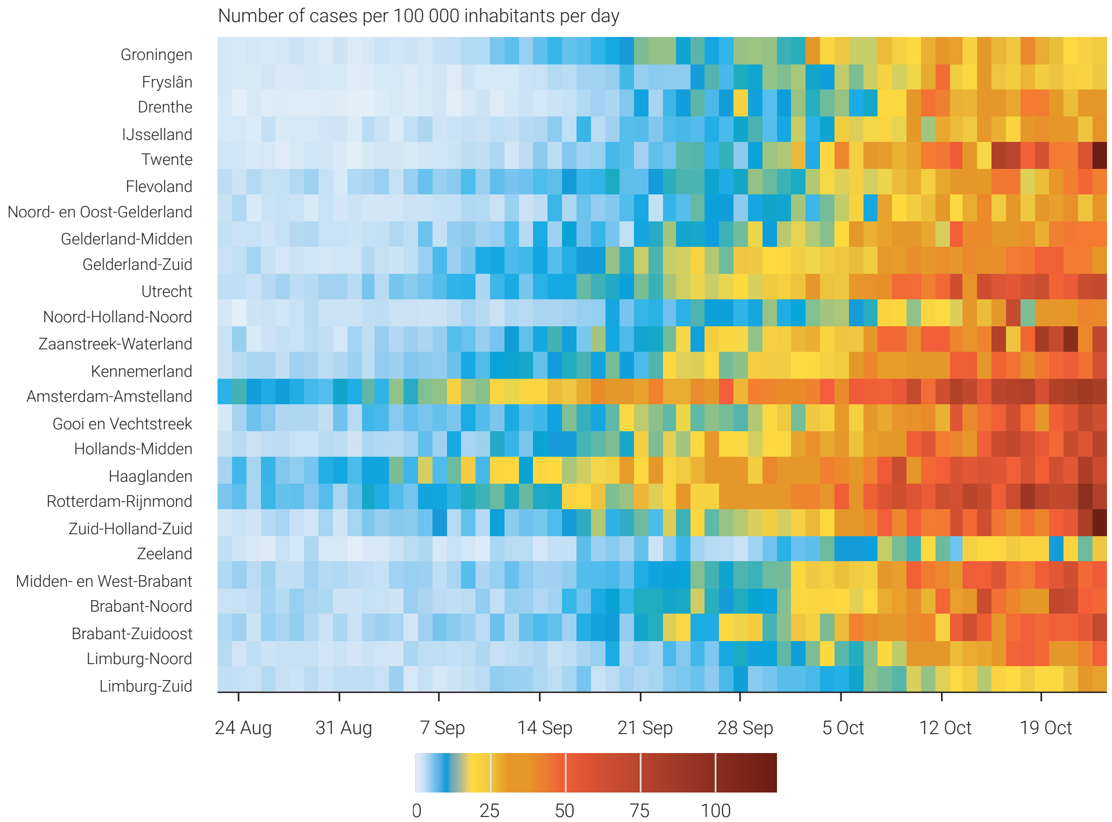
A heatmap is basically a colour-coded matrix. The name suggests that it is a map similar to those discussed in section 3.13 but it is not. In fact, it is a matrix in which a formula is used to colour each cell. The colour shade represents the relative value or risk of that cell. A diverging colour scheme is usually used, as in the example of Fig. 13.

This type of visualization is helpful because colours are easier to interpret than numbers. A heatmap makes it possible to visualize lots of timelines without these lines interfering with each other and becoming spaghetti.

Fig. 13 shows 25 regions. The data could also be shown using an interactive **motion map,** as discussed in section 3.13. This would emphazize the spatial distribution, with change over time as a secondary element. In a heatmap, the greatest emphasis is placed on the time trends, with less on the spatial dimension.

Fig. 13. Example of a heatmap

Laboratory-confirmed COVID-19 cases, by Safety Region, 2020

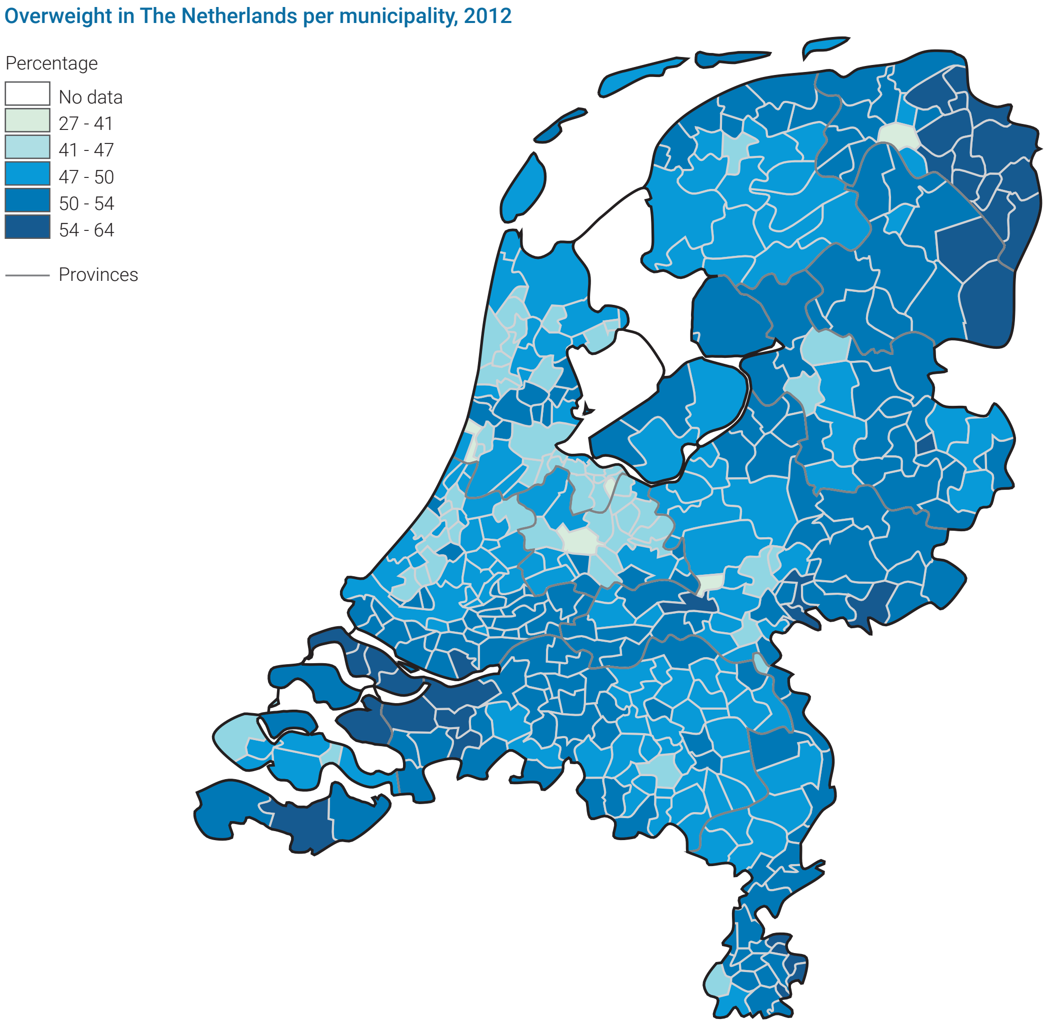


Data source: Covid-19 aantallen per gemeente per publicatiedatum [online database] [in Dutch]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2020 (https://data.rivm.nl/geonetwork/srv/dut/catalog.search#/metadata/5f6bc429-1596-490e-8618-1ed8fd768427, accessed 18 December 2020).

3.13 Map

Much of the data dealt with in epidemiology has a location element, which makes it easy to illustrate on a map. Fig. 14 is an example of a map visualization showing the percentage of people who are overweight per municipality in the Netherlands.

Fig. 14. Example of a choropleth map



Data source: Gezondheid per buurt, wijk en gemeente [online database] [in Dutch]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2020 (https://www.rivm.nl/media/smap/index.html, accessed 18 December 2020).

Two basic types of map can be distinguished:

- topographic maps used for navigation;
- thematic maps used for visualizing spatial patterns and processes.

Thematic mapping is the main instrument for visualizing statistical data. Among many different thematic maps, those that are widely used include the following:

- **Choropleth maps** show statistical data aggregated over predefined regions, such as countries or states, by colouring or shading these regions (see the example of Fig. 14).
- **Chorochromatic maps** represent a categorical or nominal variable distributed over space. Common examples include maps of surface geology, soil, vegetation, land use, city zoning and climate type.
- **Isoline maps** depict continuous quantitative fields, such as precipitation or elevation, by partitioning space into regions, each containing a consistent range of values of the field. The boundary of each region – an isoline – thus represents the set of locations of constant value.
- **Proportional symbol maps** use point symbols of different sizes (height, length, area or volume) to represent quantitative statistical values associated with different areas or locations within the map.
- **Dot density maps** place small point symbols over a given space to indicate the distribution of a given phenomenon. The location of each dot may represent the actual location of a single instance. The famous map by John Snow (see Fig. 15; *(10)*) is an example of a dot density map.
- **Flow maps** use line symbols to portray movement or relationship between two or more places, such as air travel, monetary aid or economic trade.
- **Motion maps** are interactive and can only be used in an interactive environment like a webpage or a PowerPoint presentation. Changes in time are shown like a movie. Sometimes several small maps are created, each from a different time period, and the set functions as a motion map when shown on one page.

Fig. 15. Map of the cholera outbreak in the Parish of St. James, Westminster, autumn of 1854

Source: Snow *(10)*.

The art of mapping data is called cartography. A standard reference work for cartography is Kraak and Ormeling *(11)*. Cartography is a research field in itself, and it would go beyond the scope of this guidance document to discuss all the ins and outs of maps in general and thematic maps in particular. For thematic mapping, the entry in Wikipedia is a good starting-point *(12)*.

4. The importance of side information

Every chart needs side information: it is vital to be able to understand the data shown. As noted in the introduction, side information is less important in the exploration phase, when charts are created for the analyst or the analyst's colleagues, with whom they communicate directly.

When it comes to publication of the visuals, however, side information is essential. It is also important to add and improve side information when upgrading data exploration visuals to data presentation (or explanation) visuals. A number of elements should be considered:

- title and optional subtitle of the chart
- title of the X axis
- title of the Y axis
- legend items
- source
- creator
- annotations.

4.1 Title and subtitle

The chart title is the first thing a user sees when they read a data visualization and is therefore essential. It should be easy to understand and should not just use numbers: it should use words to describe those numbers. Without an appropriate title the markers, lines and numbers could mean anything.

The title gives information on the subject a chart is about. If applicable, it should also contain the area (e.g. country, state, province) and period (e.g. year, month, time period).

Titles should be clear and concise and avoid using unnecessary words. For example, here are two titles:

- The average number of hospital admissions per day in the region of Bukhara in Uzbekistan in October 2019;
- Daily hospital admissions in Bukhara, Uzbekistan, October 2019.

The second conveys the required information with clarity and without verbiage.

If the title contains all the information, no subtitle is needed. A subtitle can be considered if the title is too long, but the title can usually just be shortened. It is also possible to use the subtitle for area and/or period indication. If done consistently, this helps guiding the user on where to look for specific information.

4.2 Title of the X axis

An X axis title should always be used unless it is entirely obvious what the X axis displays. The X axis title should contain the entity and, if applicable, the units in which the entity is measured, such as:

- weight (kg)
- length (km)
- age categories.

Often the X axis represents a time period in years or months. In such cases it is obvious that the units represent a period entity, and the X axis title can be dropped. With a nominal scale, an X axis title is not needed at all, since it would be redundant alongside either the title of the chart or the legend items.

4.3 Title of the Y axis

The title of the Y axis should always be added to the chart. It describes the units the chart displays, such as:

- number
- number (× 1 000)
- percentage
- euros (× 1 000 000)
- euros per inhabitant
- years.

It is tempting to repeat the subject of the chart in the Y axis title. Unless it is absolutely necessary, duplicating information in the chart and its side information should be avoided.

4.4 Legend items

If a chart contains only one data series a legend is not needed as it will be obvious what this data series is. If a chart contains two or more data series, however, a legend is essential. Each separate data series should be represented by a corresponding legend item.

It is advisable to keep every legend item short and sweet to avoid clutter on the chart.

4.5 Source

Information about the source of the information is extremely important for validating the quality and reliability of the charted data. For instance, if the data come from a national statistics bureau, WHO, the Organisation for Economic Co-operation and Development or another trustworthy organization, the charts have greater credibility than if data are collated from some minor source on the web. In the latter case, it is likely that results will be disputed.

4.6 Creator

It can also be helpful to mention the organization that has produced the chart. If the chart is used in another context, it is important that it contains information related to the creator. Especially on the web, charts can be used out of context of their original reports.

4.7 Annotations

Finally, there is always the option to use extra text within the plot area. For instance, if the data contain a trend break because of a different way of measuring the subject across the years, it is worth considering annotating this in the chart.

5. Best practice for making charts

5.1 Supporting the message directly

The first piece of advice is the most important: charts should not create “hidden pictures” that require the reader to extrapolate information; they should display visuals in accordance with the key message. This may seem obvious, but it can be tempting to violate this rule. The following simple example can help to explain this.

Fig. 16 shows one way to visualize the difference in life expectancy between men and women over time in the Netherlands. It supports the observation that the difference grew from 1950 and declined since 2000.

This chart, however, shows the life expectancy of men and women independently. Only indirect, derived visual observation shows the growth and decline of the difference. If the aim is to emphasize the difference, it is better to visualize the difference itself, as in Fig. 17. This displays the highest and lowest differences in life expectancy between men and women exactly. It is much clearer now to see when the growth of the difference stops and the decline starts.

Of course, Fig. 16 is a really attractive chart with useful information for a story about differences in growth trends between men and women. It would be helpful if used to show the rise in life expectancy in the Netherlands since 1950. It would also be a good foundation for a report on the way growth in life expectancy falters in different periods for men and women. To show the differences in life expectancy between men and women over time, however, Fig. 17 is the most appropriate chart.

Fig. 16. Example of a chart with a “hidden picture”

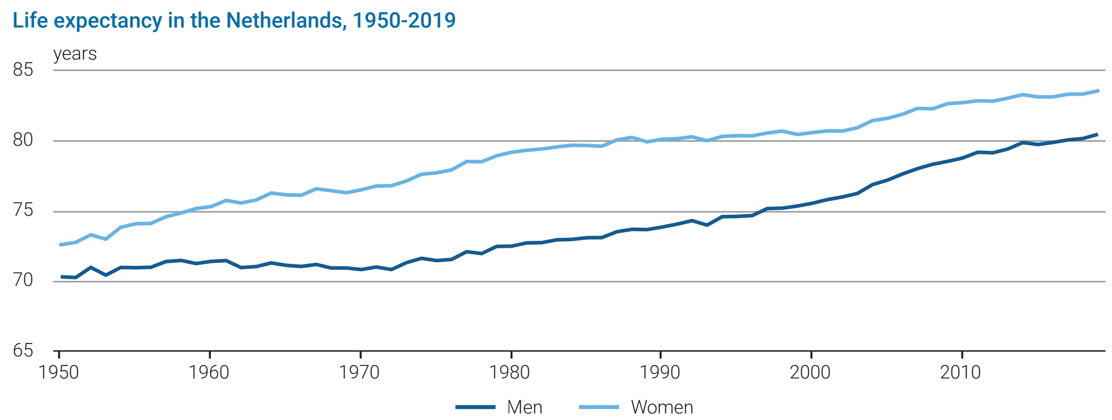
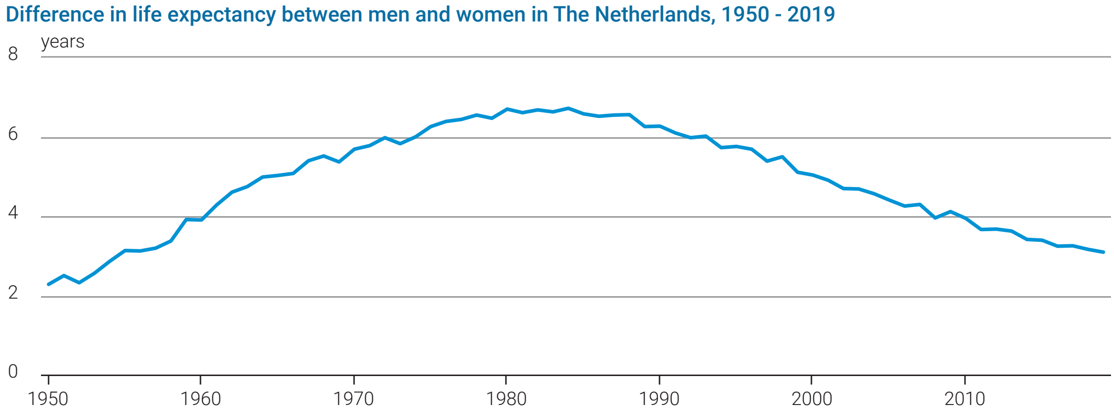


Fig. 17. Example of a chart supporting the key message



Data source: Levensverwachting; geslacht, leeftijd (per jaar en periode van vijf jaren) [online database] [in Dutch]. The Hague: Statistics Netherlands (CBS); 2020 (https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37360ned/table?dl=481ED, accessed 18 December 2020).

5.2 Avoiding chartjunk

Note that a legend is not needed in the example of Fig. 17. This leads on to the second piece of advice; avoiding "**chartjunk**". This refers to all visual elements in charts that are not required to comprehend the information represented, or that distract the viewer from this information *(13)*.

The term chartjunk was coined by Edward Tufte in *The visual display of quantitative information (9)*:

> *The interior decoration of graphics generates a lot of ink that does not tell the viewer anything new. The purpose of decoration varies – to make the graphic appear more scientific and precise, to enliven the display, to give the designer an opportunity to exercise artistic skills. Regardless of its cause, it is all non-data-ink or redundant data-ink, and it is often chartjunk.*

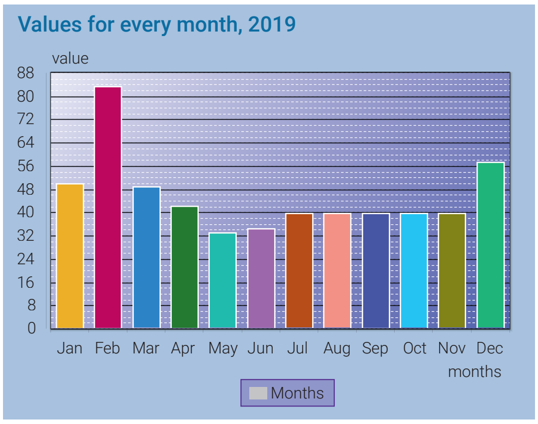
Chartjunk is closely related to the **data-ink ratio** – another term coined by Tufte. Data-ink is the non-erasable ink used for presentation of the data. If data-ink were removed from the image, the graphic would lose content. Non-data-ink is thus the ink that does not convey any information.

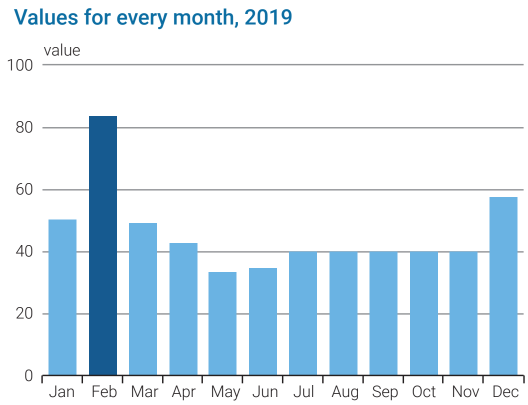
The data-ink ratio is the proportion of ink (or pixels, when displaying information on a screen) used to present actual data, without redundancy, compared to the total amount of ink (or pixels) used in the entire display, such as in a table or graph. The goal is to design a display that has the highest possible data-ink ratio (that is, as close to the total of 1.0 or 100% as possible), without eliminating something that is necessary for effective communication.

The left chart in Fig. 18 shows an example with a very low data-ink ratio. Lots of ink is used for large fonts, background colours of both the chart and the legend, the legend itself, unique colours for every datapoint and a border colour for the columns. In this case an X axis label is also unnecessary, and the gridlines are too thick and too dense. A lot of ink could be saved.

The right chart in Fig. 18 shows a clean chart with one column highlighted. This can be a useful option to support a story that says something about the highest score in February. In that case, it supports the message, and some extra data-ink can be appropriate.

Fig. 18. Example of a chartjunk chart versus a clean chart





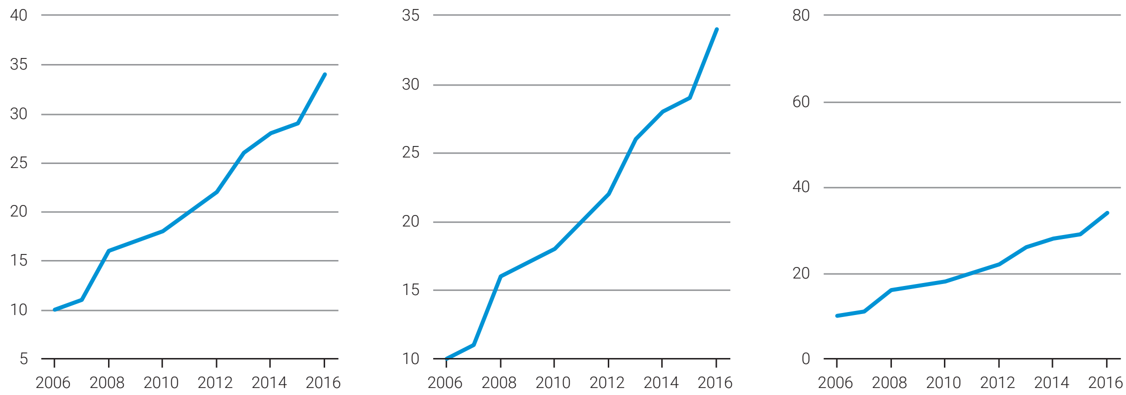
Data source: artificially generated dataset.

5.3 Considering the Y axis range and the chart aspect ratio

Note that the Y axis of a line chart does not need to start at zero (see Fig. 16). In the case of life expectancy, minor changes over time are relevant. If the Y axis started at 0, the message would be lost. On the other hand, if the Y axis is compressed too much, the message can be exaggerated. In 1954 Darryl Huff published his well known book *How to lie with statistics (14)*. This slim volume provides numerous examples that illustrate what happens if charting is done the wrong way. This section discusses some basic issues that can be avoided easily.

Playing with the range of the Y axis can be awkward. Stretching or compressing it leads to exaggeration on the one hand and trivialization on the other *(15)*. Changing the minimum and maximum points of the Y axis affects how the graph appears. Stretching the height of the graph can create fake drama, while stretching the width can underplay it. Fig. 19 shows an example.

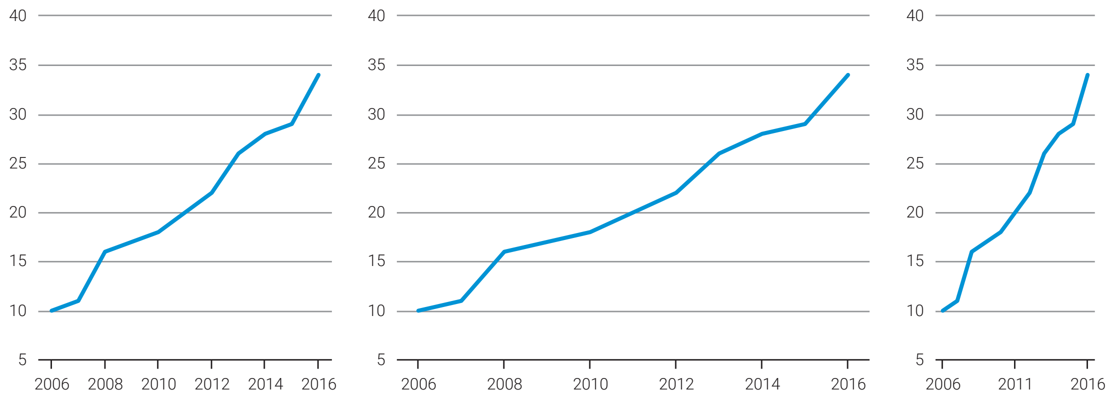
Fig. 19. Examples of a different Y axis with the same data



Data source: artificially generated dataset.

Changing the aspect ratio of a graph's dimensions also affects how the graph appears, as with Fig. 19. Fig. 20 shows the effect of different width/height ratios.

Fig. 20. Examples of changing the aspect ratio



Data source: artificially generated dataset.

As with a stretched or compressed photograph, a chart's dimensions – or its aspect ratio – can change the image being presented. While an incorrect aspect ratio in a photograph is usually immediately apparent, a distorted one in a chart can easily go unnoticed. Whether this results in an overblown or understated message, it misleads the audience.

"There is no single rule to follow in terms of how high or wide to make the graph, but a useful notion involves **banking to 45°**, whereby the average slope angle across your chart heads towards 45°," according to data visualization specialist Andy Kirk *(16)*. This may be impractical to measure, but judging by eye tends to do the trick. A good foundation for the banking to 45° rule can be found in the work of William Cleveland *(17)*.

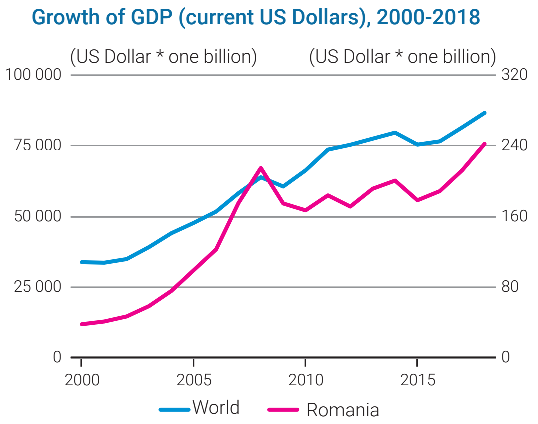
5.4 Avoiding use of two Y axes in one chart

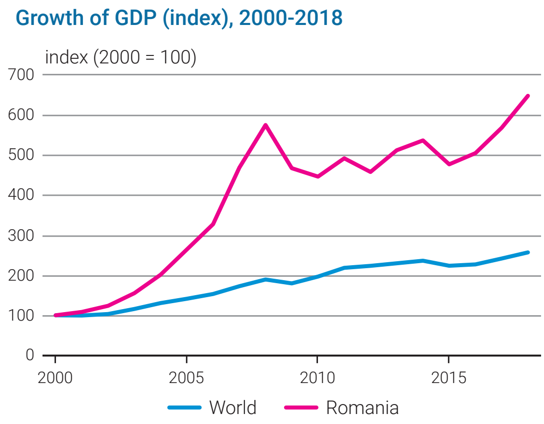
It is relatively easy and can sometimes be helpful to plot multiple sets of data using a common horizontal axis if all the data are expressed with the same unit of measurement. If they use different units, however, a secondary vertical axis on the right of the chart needs to be added. Charting two sets of data with one scale on the left and another on the right can be confusing; it can also suggest a relationship that may not exist.

This display requires time and effort from the audience to decode and understand which data should be read against which axis. Even if the reader solves this puzzle, they are tempted to compare the magnitudes of values between the two sets of data, but this is meaningless, given that the scales and units are different.

As a general rule, it is preferable to avoid squeezing too much into a small space. To tell an elaborate story, it is better to use two or more charts. If data have to be placed in one chart, indexing can be an option (see Fig. 21). A comprehensive overview of problems with two Y axes and solutions can be found in Lisa Rost's blog *(18)*.

Fig. 21. Examples of two data sets in one chart





Data source: Source: The World Bank

5.5 Avoiding 3D charts

Tools like Microsoft Excel allow 3D plots to be created easily, and they are therefore popular. The problem is that the projection of 3D objects into two dimensions for printing or display on a monitor distorts the data. The human visual system tries to correct for this distortion as it maps the 2D projection of a 3D image back into a 3D space, but this correction can only ever be partial.

Since the perception of the data in a 3D chart is always distorted, it is best to avoid them. Chapter 26 of Claus Wilke's book *Fundamentals of data visualization (19)* shows good examples of distorted data visualizations using 3D.

5.6 Avoiding pie charts

As discussed in section 3.10, a pie chart is rarely the best option. Pie charts make it more difficult for decision-makers to understand the messages they contain. They seem friendly, but in reality they are hard to read, and in most cases a better alternative is available.

Children learn fractions by imagining equal slices of a pie. But when the slices are not equal – as is often the case with real-world data – it is difficult to envisage the specific parts of a whole the pie chart is intended to convey. Human brains cannot make accurate estimates or comparisons of angles. When the slices are fairly close in size, it is difficult – if not impossible – to tell which is bigger; when they are not close in size, it may be easy to determine which is bigger, but it is not possible to judge by how much.

A bar chart is a much better choice for the job. Humans are just naturally not particularly good at distinguishing differences in slices of a circle, especially at a glance; they are much better equipped to notice differences in rectangular shapes. Further, it is also almost impossible to compare similar slices from two different pie charts. For this purpose, slope charts are best.

A final argument not to use a pie chart is from a design point of view: a pie chart takes up far more space to convey a set of data than other options. In addition, the labels do not line up, so the result can become cluttered and hard to read. Delivering precise numbers with a pie chart requires a lot of effort, such as relying on direct slice labels that might not fit or legends that make the reader’s eyes jump back and forth between the pie and the legend.

6. Creating unity in chart layouts

Another important issue is unity in chart layouts. This can make a big difference between a clear report and one that looks cluttered. The latter can draw attention away from the real message, so it is particularly important to make use of (or even create) a data visualization style guide *(20)*.

Unity lies in small but important things, including the following.

- Every chart should have a uniform font and font colour.
- A hierarchy in font size should be used for different elements of the side information. By its nature, the title of a chart is the most important element; this is followed by the subtitle, the titles of the X and Y axes, the labels placed on both axes and finally the source.
- The same margins should be used around the charts and the different elements the charts contain.
- Placement of titles and labels should be uniform.
- The way the X axis, Y axis, tick points on both axes and gridlines are drawn should be unified if possible. The same colour, width, length and style should be used for the same elements in all charts.
- Colour use should be unified. It is important to note that brand colours do not normally work for data visualization *(21)*.
- The elements in charts should be unified (line thickness in line charts, border colour and border width of columns and bars).

In this guidance document all these elements are unified in the charting examples created above.

7. Further reading

For extra reading on this topic, a great deal of content can be found on the web. In fact, so many articles, blogs, books and references are available it can be overwhelming, so this section sums up some good starting-points for further reading.

Books

Most books are not available free of charge, but there are some exceptions.

- A helpful book, with good examples and great for free further reading is: Wilke CO. Fundamentals of data visualization: a primer on making informative and compelling figures. Sebastopol, CA: O'Reilly Media; 2019 (https://clauswilke.com/dataviz/).
- A slightly older book (last updated in 2012) but still worth mentioning is: Principles of epidemiology in public health practice. Atlanta, GA: Centers for Disease Control and Prevention; 2012 (https://www.cdc.gov/careerpaths/k12teacherroadmap/classroom/principlesofepi.html).
  In this, lesson four is about displaying public health data; this chapter can be helpful for answering extra questions.

Blogs

Many blogs can be found on the web, offering good inspiration.

- Nightingale is the journal of the Data Visualization Society. It contains useful blogs and articles about data visualization, some of which are immensely helpful for giving more insight and inspiration: https://medium.com/nightingale
- Chartable is a blog written by people who work on the online charting and mapping tool Datawrapper. Like Nightingale, this contains inspirational articles about data visualization: https://blog.datawrapper.de/
- Perceptual Edge is the official blog from Stephen Few, who is well known for his ideas about data visualization. Since 2006 he has published numerous articles about a wide palette of subjects around data visualization: http://www.perceptualedge.com/

References [2]

1. Few S. What is data visualization? Perceptual Edge [blog]. 4 May 2017 (https://www.perceptualedge.com/blog/?p=2636).

2. What is data visualization and why is it important? [website]. Campbell, CA: Import.io; 2019 (https://www.import.io/post/what-is-data-visualization/).

3. Anscombe FJ. Graphs in statistical analysis. Am Stat. 1973;27(1):17–21. doi:10.1080/00031305.1973.10478966.

4. Stevens SS. On the theory of scales of measurement. Science. 1946; 103(2684):677–80 (https://psychology.okstate.edu/faculty/jgrice/psyc3214/Stevens_FourScales_1946.pdf).

5. Pearson K. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philos Trans R Soc Lond A. 1895;186:343–414. (https://bayes.wustl.edu/Manual/Pearson_1895.pdf).

6. Histogram. In: Wikipedia, the free encyclopedia [website]. Wikipedia; 2020 (https://en.wikipedia.org/wiki/Histogram).

7. Heinz G, Peterson LJ, Johnson RW, Kerk CJ. Exploring relationships in body dimensions. J Stat Educ. 2003;11(2). doi:10.1080/10691898.2003.11910711.

8. Few S. Save the pies for dessert. Perceptual Edge Newsletter. August 2007. (https://www.perceptualedge.com/articles/visual_business_intelligence/save_the_pies_for_dessert.pdf).

9. Tufte ER. The visual display of quantitative information. Cheshire, CT: Graphics Press; 1983.

10. Snow J. Map on the cholera outbreak in the Parish of St. James, Westminster, during the autumn of 1854. London; 1854 (https://commons.wikimedia.org/wiki/File:A_map_taken_from_a_report_by_Dr._John_Snow_Wellcome_L0072917.jpg; Creative Commons CC-BY 4.0 license).

11. Kraak MJ, Ormeling F. Cartography: visualization of geospatial data, fourth edition. Boca Raton, FL: CRC Press; 2020.

12. Thematic map. In: Wikipedia, the free encyclopedia [website]. Wikipedia; 2020. (https://en.wikipedia.org/wiki/Thematic_map).

2 All URLs accessed 19 November 2020.

13. Chartjunk. In: Wikipedia, the free encyclopedia [website]. Wikipedia; 2020 (https://en.wikipedia.org/wiki/Chartjunk).

14. Huff D. How to lie with statistics. New York: WW Norton & Company; 1993 [1954].

15. Misleading graph. In: Wikipedia, the free encyclopedia [website]. Wikipedia; 2020 (https://en.wikipedia.org/wiki/Misleading_graph).

16. Kirk A. Data visualization: a handbook for data driven design. London: Sage; 2016.

17. Cleveland WS, McGill ME, McGill R. The shape parameter of a two-variable graph. J Am Stat Assoc. 1988;83(402):289–300. doi:10.2307/2288843.

18. Rost LC. Why not to use two axes, and what to use instead. Chartable [blog]. 8 May 2018 (https://blog.datawrapper.de/dualaxis/).

19. Wilke CO. Fundamentals of data visualization: a primer on making informative and compelling figures. Sebastopol, CA: O'Reilly Media; 2019 (https://clauswilke.com/dataviz/).

20. Cesal A. What are data visualization style guidelines? Nightingale [blog]. 10 July 2019 (https://medium.com/nightingale/style-guidelines-92ebe166addc).

21. Cesal A. How to create brand colors for data visualization style guidelines: your brand colors don't work for data visualization. Nightingale [blog]. 13 July 2020 (https://medium.com/nightingale/how-to-create-brand-colors-for-data-visualization-style-guidelines-dbd69c586dd9).

The WHO Regional Office for Europe

The World Health Organization (WHO) is a specialized agency of the United Nations created in 1948 with the primary responsibility for international health matters and public health. The WHO Regional Office for Europe is one of six regional offices throughout the world, each with its own programme geared to the particular health conditions of the countries it serves.

Member States

Albania
Andorra
Armenia
Austria
Azerbaijan
Belarus
Belgium
Bosnia and Herzegovina
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Georgia
Germany
Greece
Hungary
Iceland
Ireland
Israel
Italy
Kazakhstan
Kyrgyzstan
Latvia
Lithuania
Luxembourg
Malta
Monaco
Montenegro
Netherlands
North Macedonia
Norway
Poland
Portugal
Republic of Moldova
Romania
Russian Federation
San Marino
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Tajikistan
Turkey
Turkmenistan
Ukraine
United Kingdom
Uzbekistan

World Health Organization
Regional Office for Europe
UN City, Marmorvej 51
DK-2100, Copenhagen Ø, Denmark
Tel: +45 45 33 70 00
Fax: +45 45 33 70 01
Email: eurocontact@who.int
Website: www.euro.who.int