FUNDAMENTALS OF DATA VISUALIZATION

Chapter 1: INTRODUCTION

*"Data visualization is part **art** and part **science**. The challenge is to get the art right without getting the science wrong, and vice versa."*

This book attempts to cover the **key principles**, **methods**, and **concepts** required to visualize data for publications, reports, or presentations

The book is divided into **three parts**.

- The first, "**From Data to Visualization**," describes different types of plots and charts, such as bar graphs, scatterplots, and pie charts. Its primary *emphasis is on the science of visualization.*

=> **Group visualizations** by the **type of message** they convey rather than by the type of data being visualized.

- The second part, "**Principles of Figure Design**," discusses various **design issues** that arise when assembling data visualizations.

=> Make **aesthetic choices** about the visual elements (*colors*, *symbols*, and *font sizes*). These choices can affect both how clear a visualization is and how elegant it looks.

- The third part, "**Miscellaneous Topics**," discusses *file formats* commonly used to store images and plots, provides thoughts about the *choice of visualization software*, and explains how to *place individual figures* into the context of a larger document.

Ugly, Bad, and Wrong Figures

Ugly: A figure that has **aesthetic problems** but otherwise is clear and informative.

Bad: A figure that has **problems** related to **perception**; it may be unclear, confusing, overly complicated, or deceiving.

Wrong: A figure that has **problems** related to **mathematics**; it is objectively incorrect

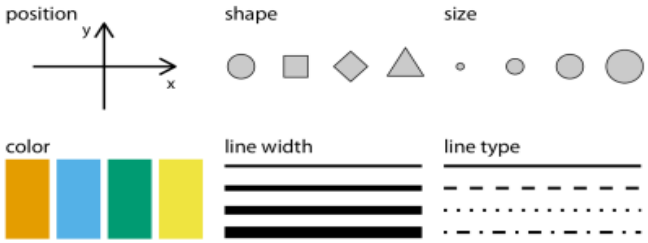
PART I: FROM DATA TO VISUALIZATION

Chapter 2: VISUALIZING DATA: MAPPING DATA ONTO AESTHETICS

*"The key insight is the following: all data visualizations map **data values** into **quantifiable features** of the resulting graphic. We refer to these features as **aesthetics**."*

Aesthetics and Types of Data

- A critical component of every graphical element is of course its **position** - where the element is located
- Next, all graphical elements have a **shape**, a **size**, and a **color.**
- Finally, if we are using lines to visualize data, these lines may have different **widths** or **dash–dot patterns**



*Some of these aesthetics can represent both **continuous** and **discrete** data (position, size, line width, color), while others can usually only represent discrete data (shape, line type)*

- **Continuous**: Can divide infinitely (50.3222 s, 132.32 cm,...)
- **Discrete**: Cannot (number of person,...)

Table 2-1. Types of variables encountered in typical data visualization scenarios.

Type of variable	Examples	Appropriate scale	Description
Quantitative/ numerical continuous	1.3, 5.7, 83, 1.5×10^{-2}	Continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.
Quantitative/ numerical discrete	1, 2, 3, 4	Discrete	Numbers in discrete units. These are most commonly but not necessarily integers. For example, the numbers 0.5, 1.0, 1.5 could also be treated as discrete if intermediate values cannot exist in the given dataset.
Qualitative/ categorical unordered	dog, cat, fish	Discrete	Categories without order. These are discrete and unique categories that have no inherent order. These variables are also called *factors*.
Qualitative/ categorical ordered	good, fair, poor	Discrete	Categories with order. These are discrete and unique categories with an order. For example, "fair" always lies between "good" and "poor." These variables are also called *ordered factors*.
Date or time	Jan. 5 2018, 8:03am	Continuous or discrete	Specific days and/or times. Also generic dates, such as July 4 or Dec. 25 (without year).
Text	The quick brown fox jumps over the lazy dog.	None, or discrete	Free-form text. Can be treated as categorical if needed.

Scales Map Data Values onto Aesthetics

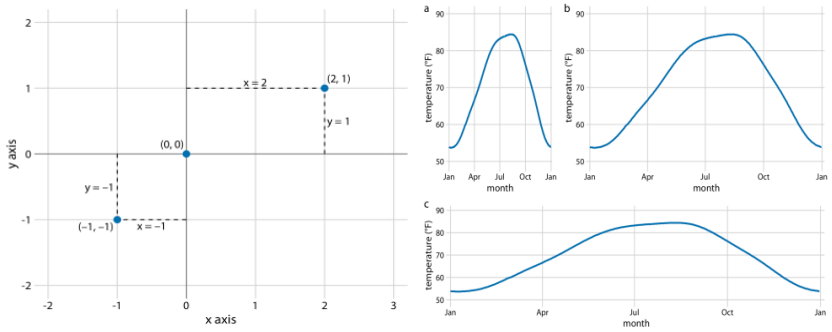
*We may need to specify which data values are represented by particular **shapes** or **colors**. This mapping between **data values** and **aesthetics values** is created via **scales**.*

Chapter 3: COORDINATE SYSTEMS AND AXES

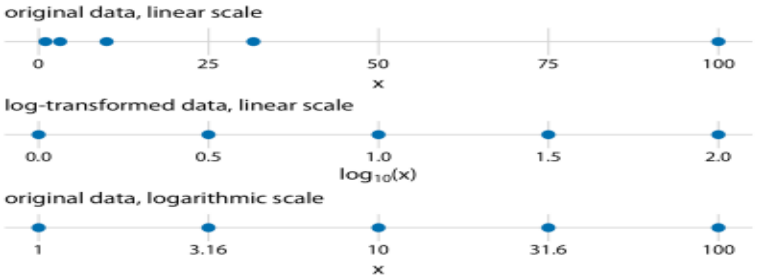
*"The combination of a set of **position scales** and their relative **geometric arrangement** is called a **coordinate system."***

Cartesian Coordinates

=> The most widely used coordinate system for data visualization.



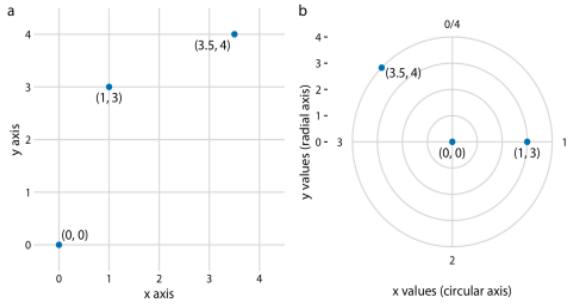
Nonlinear Axes



A nonlinear axis, or nonuniform axis, is useful in two major scenarios:

- **Extreme** data points --> misleading charts. A nonlinear axis can help better represent values that span several orders of magnitude.
- In the fields of **finance** or **science**, such as *compound interest*, *growth rate*, *earthquake strength, sound loudness, and light intensity*. By switching to a nonlinear axis, equal percentage changes can be represented as the same linear distance on a scale.

Coordinate Systems with Curved Axes



=> Apply in Navigation to indicate direction,

Chapter 4: COLORS SCALE

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

- **Distinguish** groups of data from each other
- **Represent data values**
- **Highlight.**

The types of colors we use and the way in which we use them are quite different for these three cases.

Color as a Tool to Distinguish

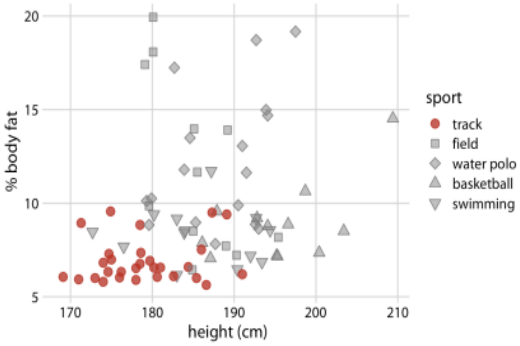
Use a **qualitative** color scale. Such a scale contains a **finite set of specific colors** that are chosen to *look clearly distinct* while also being *equivalent* to each other. The second condition requires that *no one color should stand out* relative to the others. Also, the colors *should not create the impression of an order*

Represent data values

Use a **sequential** color scale. Such a scale contains a **sequence of colors** that *clearly indicate which values are larger or smaller* than which other ones, and *how distant two specific values are* from each other. The second point implies that the color scale needs to be perceived to vary uniformly across its entire range.

Color as a Tool to Highlight

Color can also be an effective tool to **highlight specific elements** in the data. There may be specific categories or values in the dataset that carry **key information** about the story we want to tell, and we can strengthen the story by emphasizing the relevant figure elements.

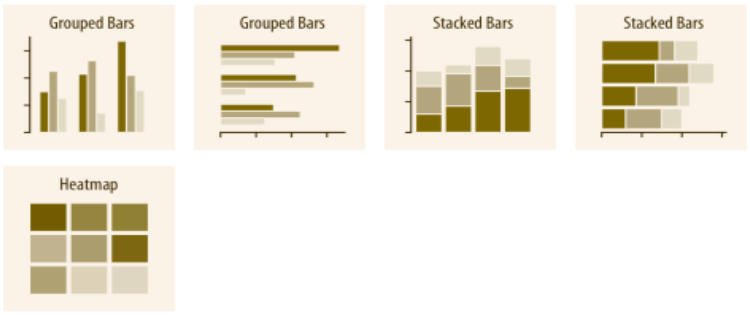


Chapter 5: DIRECTORY OF VISUALIZATION

"This chapter provides a quick visual overview of the various plots and charts that are commonly used to visualize different types of data."

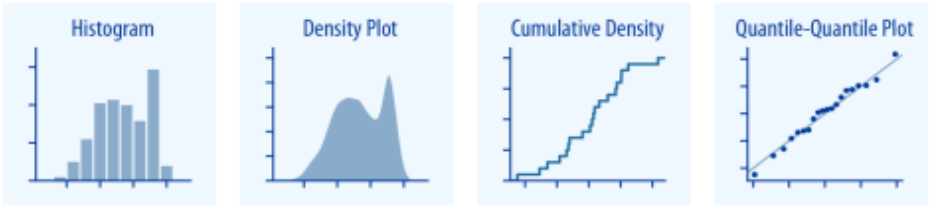
Amounts

The most common approach to visualizing amounts (numerical values for set of categories) is using **bars**, (vertically or horizontally). However, instead of using bars, we can also **place dots** at the location where the corresponding bar would end.



If there are two or more sets of categories - we can *group or stack the bars*. We can also map the categories onto the x and y axes and *show amounts by color*, via a heatmap.

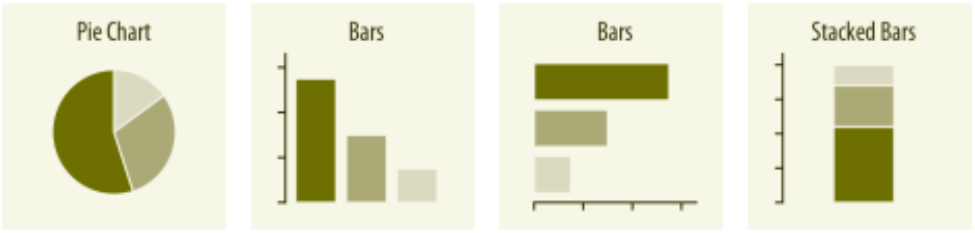
Distributions



Histograms and **density plots** provide *the most intuitive visualizations of a distribution*, but both require arbitrary parameter choices and can be misleading. **Cumulative densities** and **quantile-quantile** (q-q) plots *always represent the data faithfully* but can be more difficult to interpret.

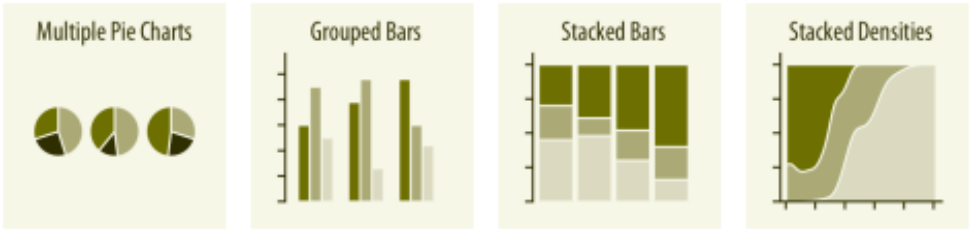
Boxplots, violin plots, strip charts, and **sina plots** are useful when we want to *visualize many distributions at once* and/or if we are primarily interested in *overall shifts among the distributions.* **Stacked histograms** and **overlapping densities** allow a *more in-depth comparison of a smaller number of distributions.* **Ridgeline** is useful when *visualizing very large numbers of distributions* or *changes in distributions over time.*

Proportions



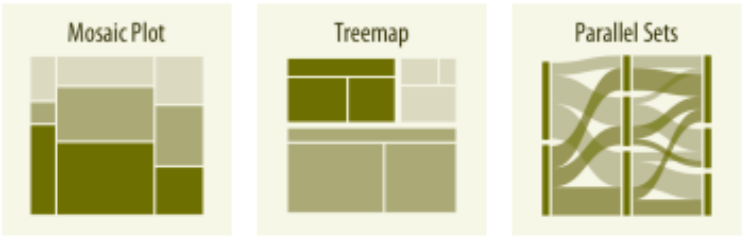
Proportions can be visualized as **pie charts, side-by-side bars,** or **stacked bars**.

- **Bars chart**: the bars can be arranged either vertically or horizontally.
- **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.*
- **Stacked bars** look awkward for a single set of proportions but can be useful when *comparing multiple sets of proportions.*



When visualizing **multiple sets of proportions** or **changes in proportions across conditions**

- **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.*
- **Stacked bars** can work for large numbers of conditions.
- **Stacked densities** are appropriate when *the proportions change along a continuous variable.*

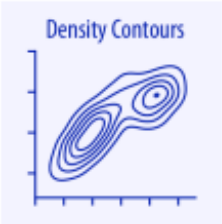


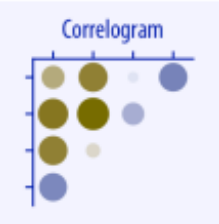
When **proportions are specified according to multiple grouping variables**:

- **Mosaic plots:** assume that every level of one grouping variable can be combined with every level of another grouping variable
- **Treemaps:** do not make such an assumption. Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.
- **Parallel** sets work better than either mosaic plots or treemaps when there are more than two grouping variables

x–y relationships

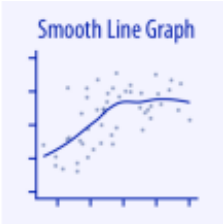
- **Scatterplots:** represent the archetypical visualization when we want to show one quantitative variable relative to another.
- If we have three quantitative variables, we can map one onto the dot size, creating a variant of the scatterplot called a **bubble chart**.
- For paired data, where the variables along the x and y axes are measured in the same units, it is generally helpful to add a line indicating x = y.
- Paired data can also be shown as a **slopegraph** of paired points connected by straight lines.





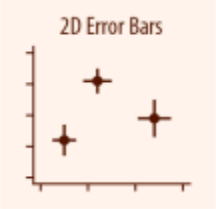
For large numbers of points, regular scatterplots can become uninformative due to overplotting.

When we want to visualize more than two quantities, on the other hand, we may choose to plot correlation coefficients in the form of a correlogram instead of the underlying raw data.

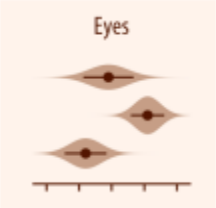


- When the x axis represents time or a strictly increasing quantity such as a treatment dose, we commonly draw **line graphs**.
- If we have a temporal sequence of two response variables we can draw a **connected scatterplot**, where we first plot the two response variables in a scatterplot and then connect dots corresponding to adjacent time points.
- We can use **smooth lines** to represent trends in a larger dataset

Uncertainty

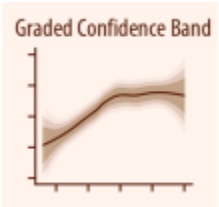


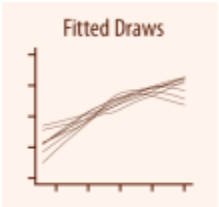
- **Error bars** are meant to indicate the range of likely values for some estimate or measurement. They extend horizontally and/or vertically from some reference point representing the estimate or measurement. Reference points can be shown in various ways, such as by dots or by bars.
- **Graded error bars** show multiple ranges at the same time, where each range corresponds to a different degree of confidence. They are in effect multiple error bars with different line thicknesses plotted on top of each other.



To achieve a more detailed visualization than is possible with error bars or graded error bars, we can visualize the actual confidence or posterior distributions.

- **Confidence strips** provide a visual sense of uncertainty but are difficult to read accurately.
- **Eyes** and **half-eyes** combine error bars with approaches to visualize distributions (violins and ridgelines, respectively), and thus show both precise ranges for some confidence levels and the overall uncertainty distribution.
- **A quantile dot plot** can serve as an alternative visualization of an uncertainty distribution. Because it shows the distribution in discrete units, the quantile dot plot is not as precise but can be easier to read than the continuous distribution shown by a violin or ridgeline plot





- For smooth line graphs, the equivalent of an error bar is a **confidence band**. It shows a range of values the line might pass through at a given confidence level.
- Like with error bars, we can draw **graded confidence bands** that show multiple confidence levels at once.
- We can also show individual **fitted draws** in lieu of or in addition to the confidence bands.

Chap 6 – Chap 16: Detail about each type of chart

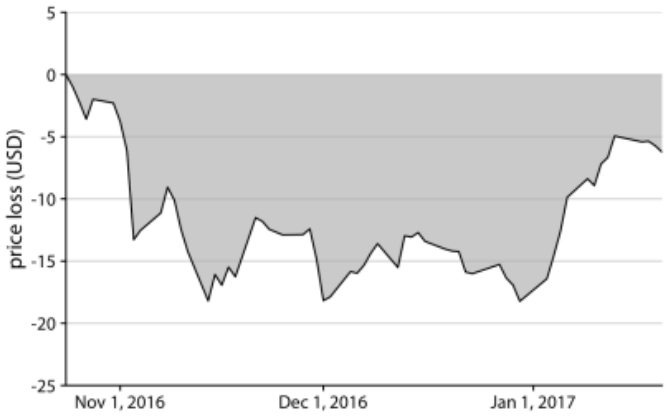
PART II: FROM DATA TO VISUALIZATION

Chapter 17: THE PRINCIPLE OF PROPORTIONAL INK

Visualizations Along Linear Axes

Bars on a linear scale should always **start at 0.**

We can draw the change over time as the difference from its first value. By **shading an area** that represents the **distance** from the high point, we are accurately representing the absolute magnitude of the change.



Visualizations Along Logarithmic Axes

Direct Area Visualizations

*That human perception is **better at judging distances** than at judging areas*

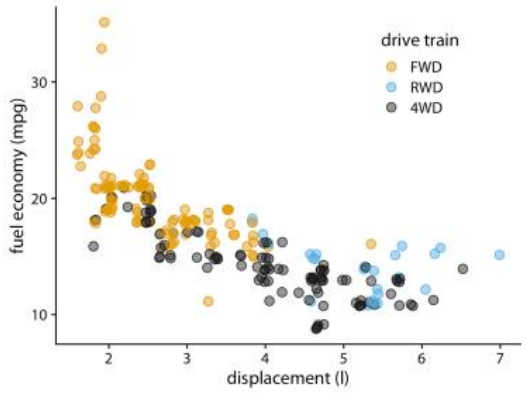
Chapter 18: HANDLING OVERLLAPPING POINTS

When visualize **large datasets**, even in small datasets if data values were recorded with low precision or rounded

=> x–y scatterplots do not work very well because many points lie on top of each other and partially or fully overlap.

=> The technical term commonly used to describe this situation is **overplotting**, which means that we are plotting many points on top of each other

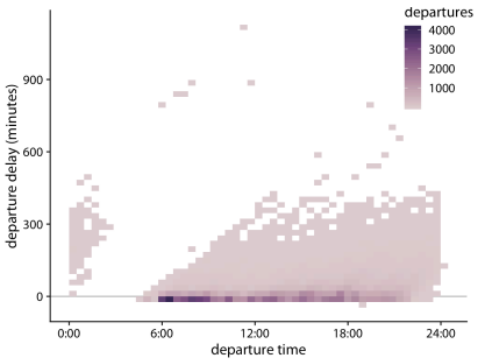
Partial Transparency and Jittering



- **Partial transparency** makes individual points **partially transparent**, so that overplotted points appear as darker points. This can help to show the density of points in a particular area of the graph. However, it can be *difficult to estimate how many points are overlapping,* and it may not be clear to the viewer that the darker points are actually multiple points.
- **Jittering** randomly displaces each point by a small amount in either the x or y direction (or both). This can help to separate overlapping points and make them easier to see. However, it is important to use jittering sparingly, as too much jitter can distort the data and make it difficult to interpret.

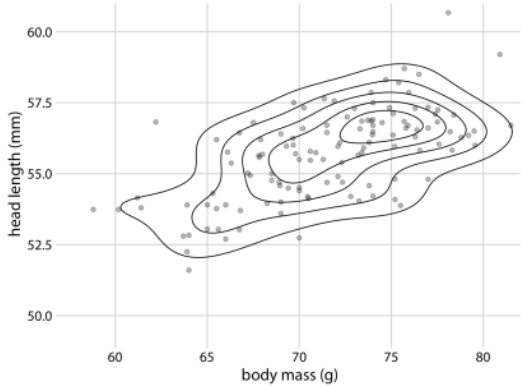
Technique	Pros	Cons
Partial transparency	Shows the density of points	Can be difficult to estimate how many points are overlapping
Jittering	Separates overlapping points	Can distort the data if used too much

2D Histograms



- 2D histograms can be used to visualize data with a **large number of overlapping points.**
- 2D histograms divide the data into two dimensions and count the number of points in each bin.
- 2D histograms can be used to **identify patterns** in the data.
- **Hexagons** are a better choice than rectangles for binning data because the points are, on average, closer to the center of the hexagon.

Contour Lines



- Visualizing **elevation**: Contour lines are often used to visualize elevation data. This is because they can be used to show the **shape of a terrain**, such as a mountain range or a valley.
- Visualizing **temperature**: Contour lines can also be used to visualize temperature data. This is because they can be used to show the **distribution of temperature** over a region.
- Visualizing **pressure**: Contour lines can also be used to visualize pressure data. This is because they can be used to show the distribution of pressure over a region.
- Visualizing other data: Contour lines can also be used to visualize other types of data, such as **population density or rainfall.**

Chapter 19: COMMON PITFALLS OF COLOR USE

*"Color can be an incredibly effective tool to **enhance data visualizations**. At the same time, poor color choices can ruin an otherwise excellent visualization. Color needs to be applied to serve a purpose, it **must be clear**, and it **must not distract**."*

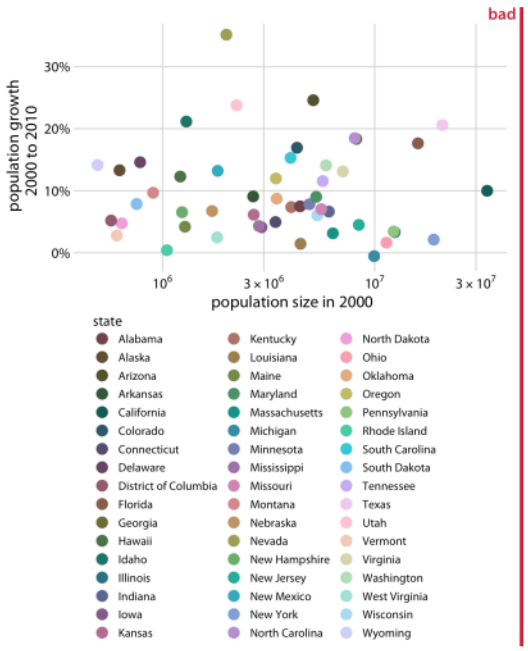
Encoding Too Much or Irrelevant Information

*One common mistake is trying to give color a job that is too big for it to handle, by **encoding too many different items in different colors.***

As a rule of thumb, qualitative color scales work best when there are **three to five** different categories that need to be colored.

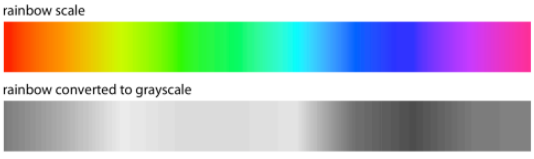
=> Use **direct labeling** instead of colors when you need to distinguish between **more than about eight** categorical items.

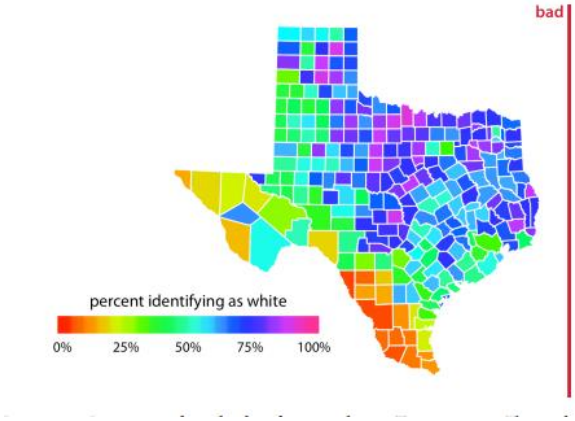
Avoid large filled areas of overly **saturated colors (màu bão hòa)**. They make it difficult for your reader to carefully inspect your figure.



Using Nonmonotonic Color Scales to Encode Data Values

Nonmonotonic: It has regions where *colors change very slowly* and others where *colors change rapidly*





Not Designing for Color-Vision Deficiency

Whenever we are choosing colors for a visualization, we need to keep in mind that *a good proportion** of our readers may have some form of **color-vision deficiency** (i.e., are colorblind).

*Approximately **8% of males** and **0.5% of females** suffer from some sort of color-vision deficiency (CVD).

Figure 19-7. A red–green contrast becomes indistinguishable under red–green CVD (deuteranomaly or protanomaly).

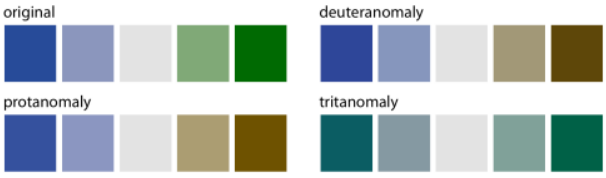


Figure 19-8. A blue–green contrast becomes indistinguishable under blue–yellow CVD (tritanomaly).

Note: Deuteranomaly, Protanomaly and tritanomaly are common types of color-vision deficiency

- Protanomaly & Deuteranomaly: cannot recognize **red-green** contrast.
- Tritanomaly: Cannot recognize **blue-green** contrast

=> To make sure your figures work for people with CVD, don't just rely on specific color scales. Instead, **test your figures in a CVD simulator.**

Chapter 20: REDUNDANT CODING

*"Encode data redundantly, using **multiple different aesthetic dimensions** instead of just using color. (When we have many different items we want to identify)"*

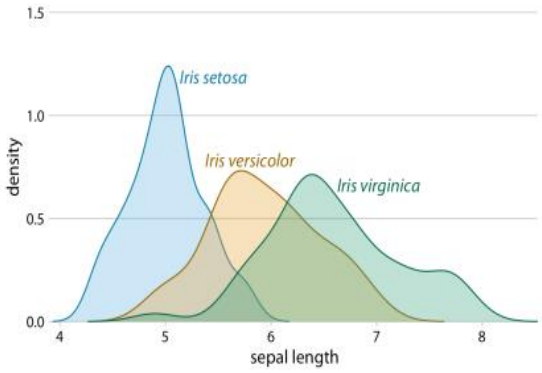
Designing Legends with Redundant Coding

- Use distinct **colors, shapes,** and **sizes** for each group of data. This will help people to easily distinguish between the different groups.
- Avoid using **colors** that are **close together** on the color wheel, such as red and orange, or green and blue. This can make it difficult for people with color blindness to distinguish between the colors.
- If you are using **line types**, *avoid using too many different types.* This can make it difficult for people to keep track of the different lines.

- Always **match the order** of the data points in the **legend** to the order of the groups in the **plot**. This will make it easier for people to associate the data points with the correct group.

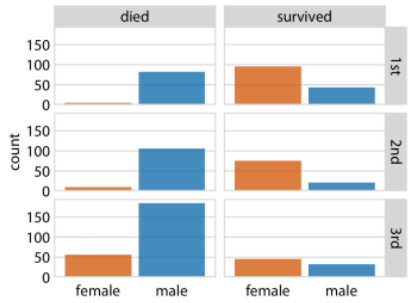
Designing Figures Without Legends

***Whenever possible**, design your figures so they don't need a separate legend*



Chapter 21: MULTIPANEL FIGURES

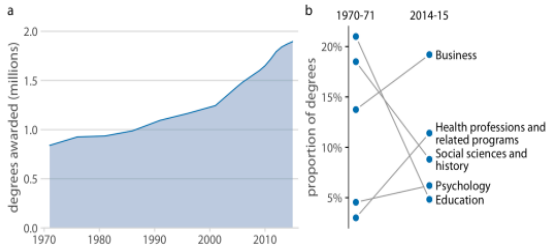
*"When datasets become large and complex, they often contain much more information than can reasonably be shown in a single figure panel. To visualize such datasets, it can be helpful to create **multipanel figures**."*



Small multiples: plots consisting of multiple panels arranged in a regular grid. Each panel shows a different subset of the data but all panels use the same type of visualization

=> Always arrange the panels in a small multiples

Compound figures: consist of separate figure panels assembled in an arbitrary arrangement (which may or may not be grid-based) and showing entirely different visualizations, or



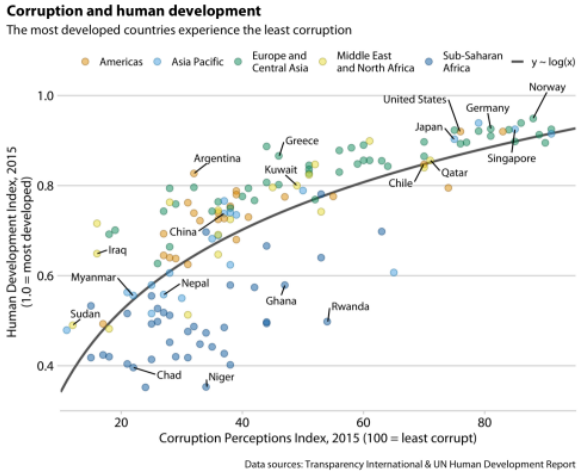
Chapter 22: TITLES, CAPTIONS AND TABLES

*"A data visualization is not a piece of art meant to be looked at only for its aesthetically pleasing features. Instead, its purpose is to **convey information** and **make a point**. To reliably achieve this goal when preparing visualizations, we have to place the data into context and provide accompanying **titles, captions,** and other **annotations**."*

Figure Titles and Captions

One critical component of every figure is the title. **Every figure needs a title.** The job of the title is to *accurately convey to the reader what the figure is about.*

A title does not have to be a complete sentence, though short sentences making a clear assertion can serve as titles.



Axis and Legend Titles

Just like every plot needs a title, axes and legends need titles as well. (Axis titles are often colloquially referred to as axis labels.) Axis and legend titles and labels explain what the **displayed data values** are and how they map to **plot aesthetics.**

Tables

Some key rules for table layout are the following:

1. Do not use vertical lines.

2. Do not use horizontal lines between data rows. (Horizontal lines as a separator between the title row and the first data row or as a frame for the entire table are fine.)

3. Text columns should be left aligned.

4. Number columns should be right aligned and should use the same number of decimal digits throughout.

5. Columns containing single characters should be centered.

6. The header fields should be aligned with their data; i.e., the heading for a text column will be left aligned and the heading for a number column will be right aligned

a ugly

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

b ugly

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

c

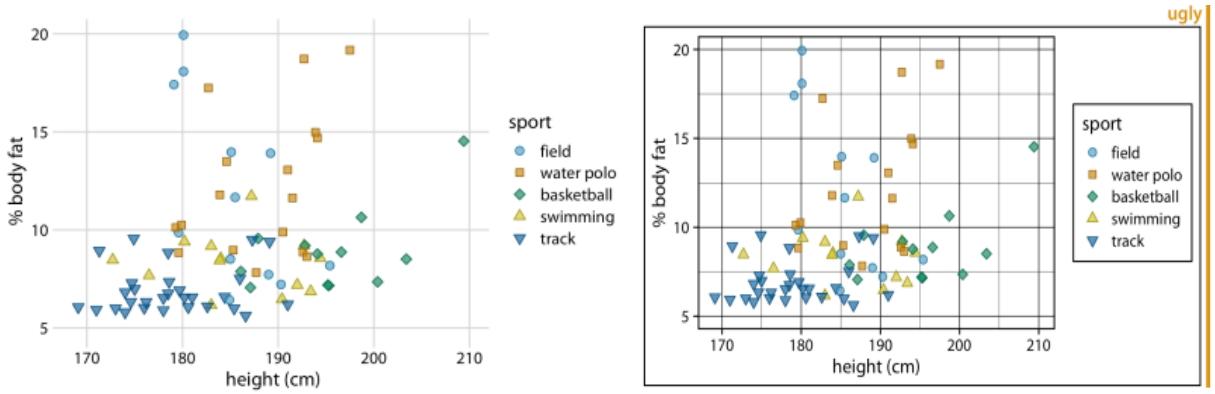
Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

d

Rank	Title	Amount
1	*Star Wars: The Last Jedi*	$71,565,498
2	*Jumanji: Welcome to the Jungle*	$36,169,328
3	*Pitch Perfect 3*	$19,928,525
4	*The Greatest Showman*	$8,805,843
5	*Ferdinand*	$7,316,746

Chapter 23: BALANCE THE DATA AND THE CONTEXT

Providing the Appropriate Amount of Context

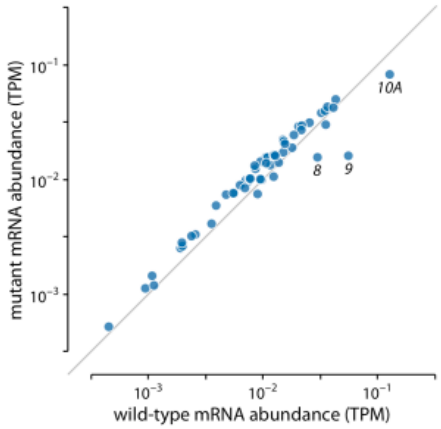


Background Grids

Grid lines in the background of a plot can help the reader discern specific data values and compare values in one part of a plot to values in another part. At the same time, grid lines can add visual noise

Paired Data

For figures where the relevant comparison is the x = y line, such as in scatterplots of paired data, I prefer to draw a diagonal line rather than a grid.



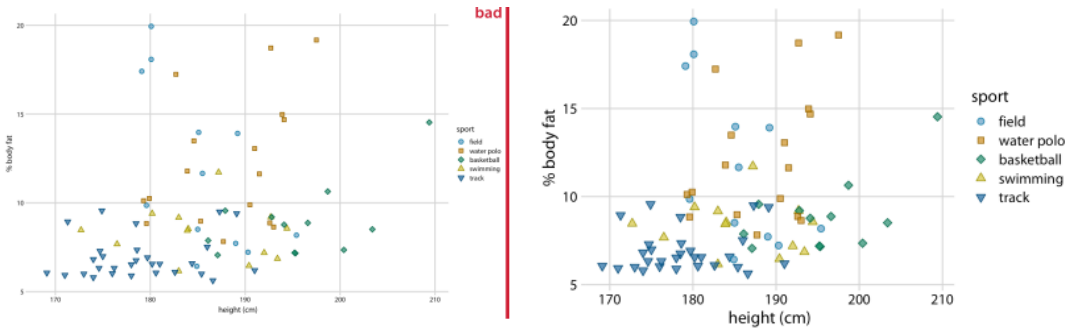
Summary

Both overloading a figure with non-data ink and excessively erasing non-data ink can result in poor figure design. We need to find a healthy medium, where the data points are the main emphasis of the figure while sufficient context is provided about what data is shown, where the points lie relative to each other, and what they mean.

Chapter 24: USE LARGER AXIS LABELS

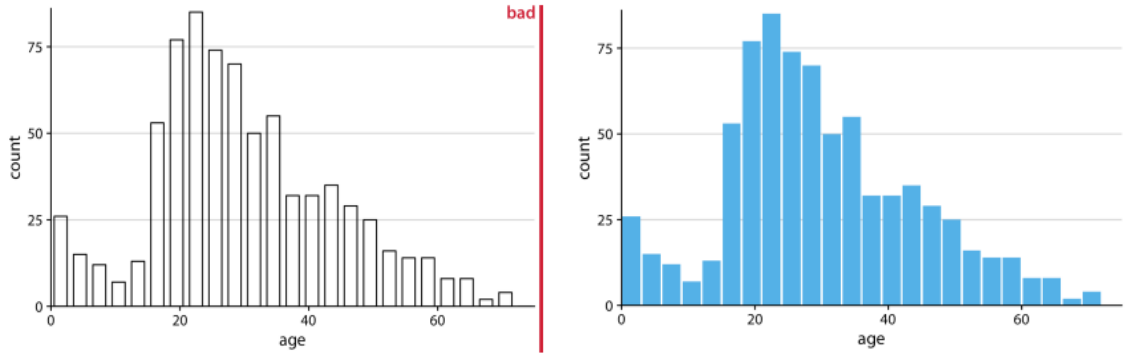
If you take away only one single lesson from this book, make it this one: pay attention to your **axis labels, axis tick labels,** and other assorted **plot annotations**. Chances are they are too small. In my experience, nearly all graphing software and plot libraries have poor defaults. If you use the default values, you're almost certainly making a poor choice

=> Always look at scaled-down versions of your figures to make sure the axis labels are appropriately sized



Chapter 25: AVOID LINE DRAWINGS

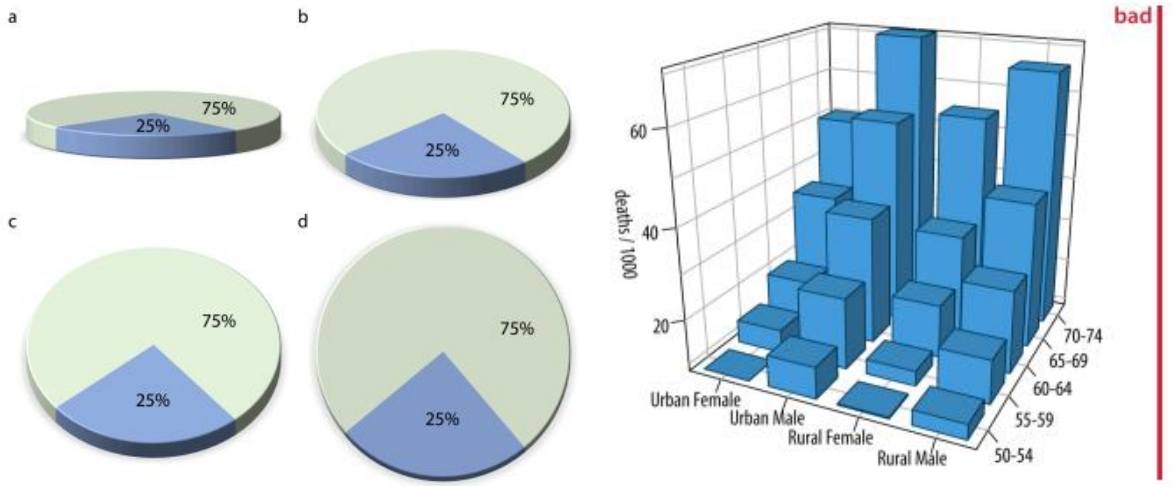
*"Whenever possible, visualize your data with **solid, colored shapes** rather than with lines that outline those shapes. Solid shapes are more **easily perceived** as coherent objects, are less likely to create visual artifacts or optical illusions, and more **immediately convey** amounts than do outlines."*



Chapter 26: DONT GO 3D

Avoid Gratuitous 3D

Avoid 3D Position Scales



PART III: MISCELLANEOUS TOPICS

Chapter 27: UNDERSTANDING THE MOST COMMONLY USED IMAGE FILE FORMAT

Chapter 28: CHOOSING THE RIGHT VISUALIZATION SOFTWARE

*"The best visualization software is the one that allows you to **make the figures you need**."*

Reproducibility and Repeatability

In **scientific experiments:**

- Work is considered **reproducible** *if the same results can be obtained by a different research group performing the same type of study.* (nhóm khác làm lại vẫn ra cùng kqua)
- Work is **repeatable** if similar or *identical measurements* can be obtained by the same person repeating the exact same measurement procedure on the same equipment. (tự mình làm lại vẫn ra cùng kqua – về mặt metric, measurement)

In **data visualization,**

- A visualization is **reproducible** if the plotted data is available and any data transformations applied before plotting are exactly specified.
- A visualization is **repeatable** if it is possible to recreate the exact same visual appearance, down to the last pixel, from the raw data.

Data Exploration Versus Data Presentation

Data exploration phase, the focus is on **understanding the dataset's key features** by trying different types of visualizations, data transformations, and subsets of the data. The speed and efficiency of iterating through different ways of looking at the data are crucial in this phase. The aesthetics of the figures created during this phase are secondary to the patterns in the data.

Data presentation phase, the goal is to prepare a **high-quality, publication-ready figure**. This phase begins once you *understand your dataset* and know what *aspects you want to show to your audience*.

Separation of Content and Design

⇨ *Use 1 tool for complicate cleaning & Other for visualize*

Chapter 29: TELLING A STORY AND MAKING A POINT

*"Most data visualization is done for the **purpose of communication**. We have an insight about a dataset, and we have a potential audience, and we would like to convey our insight to our audience."*

⇨ **If we don't provide a clear story ourselves, then our audience will make one up.**

*#All Knowledge from past to present is **fascinating**, just that they haven't been **properly told.***

What Is a Story?

A story is a set of *observations, facts, or events, true or invented*, that are presented in a *specific order* such that they create an **emotional reaction** in the audience.

⇨ The **emotional reaction** is created through the **buildup of tension** at the *beginning* of the story followed by some type of **resolution** toward the *end of the story.* (**Story Arc**)

- Opening => Challenge => Action => Resolution format

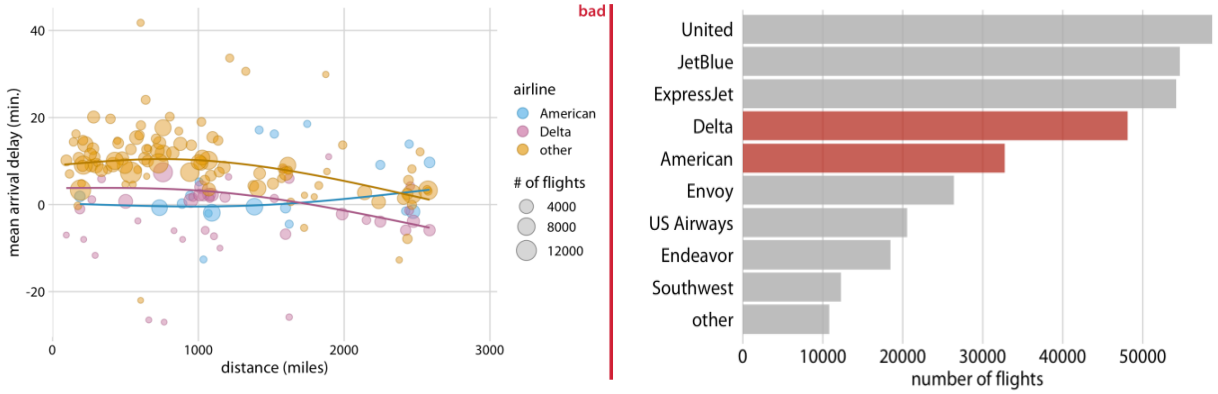
⇨ Using **multiple visualizations** to tell a complete story in data presentation

Make a Figure for the General

"Never assume** your audience **can rapidly process complex visual displays"

⇨ **Simplifying** figures to **highlight** only the **important** points, making them easily understandable to a broad audience.

⇨ **Avoid the temptation** to create overly **complex visualizations** due to the capabilities of modern visualization software. => *Don't effectively convey a meaningful story.*



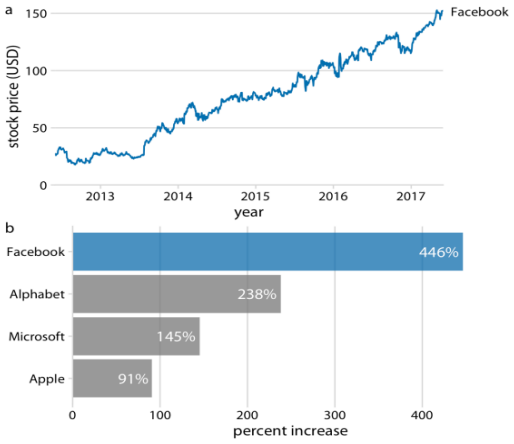
Build Up Toward Complex Figure

When we want to show more **complex figures** that contain a large amount of information at once.

- ⇨ **Make things easier** for our readers if we first show them a **simplified version** of the figure before we show the **final one** in its full complexity.

Research shows that more **visually complex** and **unique figures** are more **memorable**, but they may hinder a person's ability to get a quick overview of the information or make it difficult to distinguish small differences in values.

- ⇨ A balance between *clarity* and *memorability* is needed.



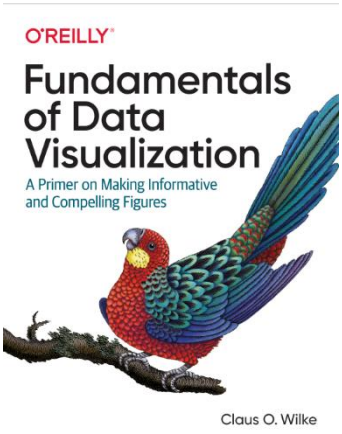
Be Consistent but Don't Be Repetitive

When preparing a presentation or report, aim to use **a different type of visualization** for each **distinct analysis.**

THE END

* The animal on the cover of Fundamentals of Data Visualization is a *western rosella parakeet* (Platycercus **icterotis**) – a small species of parrot is Australia. The name **icterotis** is derived from ancient Greek for "yellow ear," referring to the yellow spot on each of the bird's cheeks.

=> The rosella is very **colorful** indeed—it has a **red head & neck**, **a barred green**, **black**, and **red back**, **blue wing feathers & blue green tail**



Source: Fundamentals of Data Visualization – Clau O.Wilke

Summarize by Tri Hai – 2023/09/17

```
SELECT Knowledge
FROM Past_to_present        -> 'Hello World'
WHERE Properly_told = True ;
```

Mutsukkn – A story lover