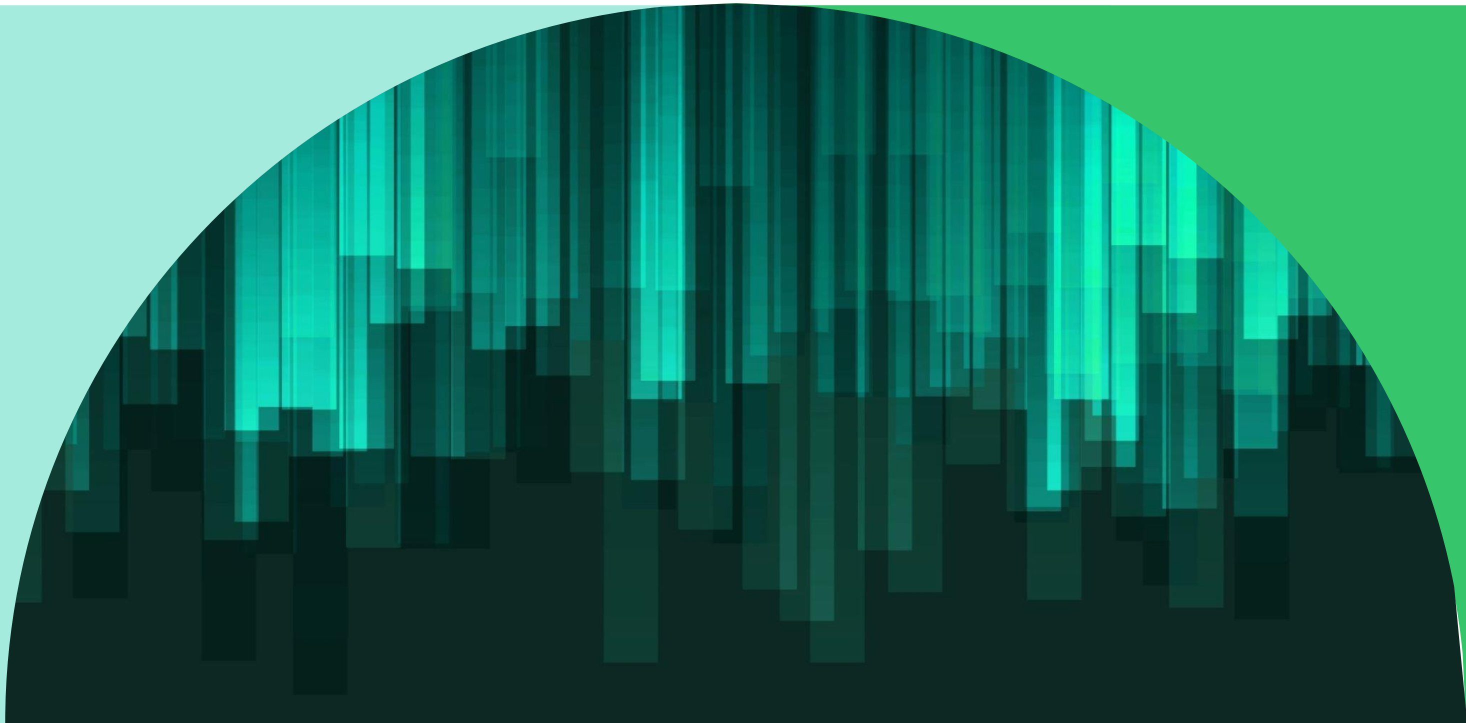
Data Visualization Principles

Jerome Niyirora, PhD
SUNY Polytechnic Institute

Reference: Fundamentals of Data Visualization by Claus O. Wilke



Aesthetics

- Whenever we visualize data, we take **data values and convert them in a systematic and logical way into the visual elements that make up the final graphic.**

- All data visualizations map data values into quantifiable features the make up the **aesthetics** of the final graphic.

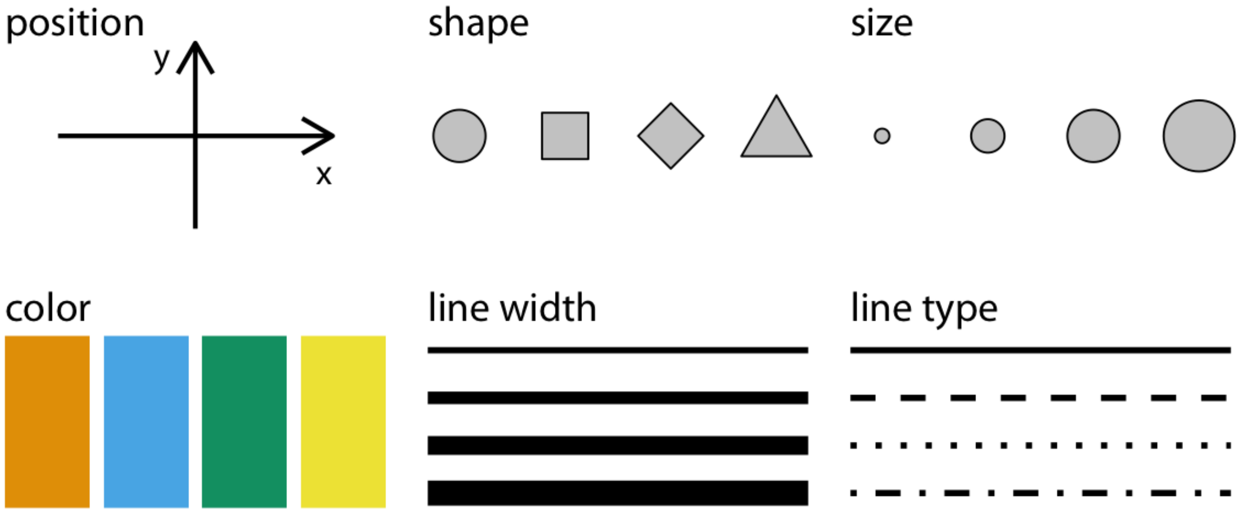
Aesthetics

- To represent continuous and discrete data

- Use graphical representations- must have a shape, a size, and a color

- Use lines – with different widths or dash-dot patterns

- For text, specify font family, font face, and font size.



Source: https://clauswilke.com/dataviz/aesthetic-mapping.html#aesthetics-and-types-of-data

A directory of visualizations

Amounts

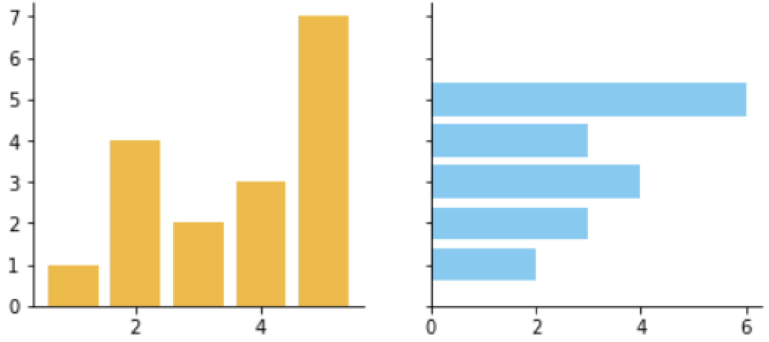
Amounts

- The most common plots for visualizing amounts include **bar charts**, **dot charts**, and **heatmaps**. The amount over time can be visualized using **line charts**.

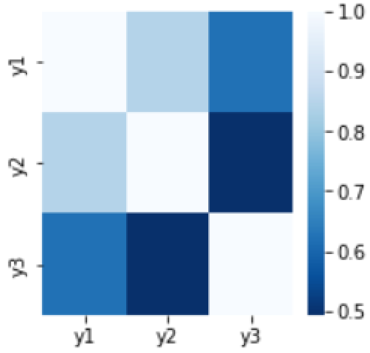
Amounts

- The most common plots for visualizing amounts include **bar charts**, **dot charts**, and **heatmaps**. The amount over time can be visualized using **line charts**.

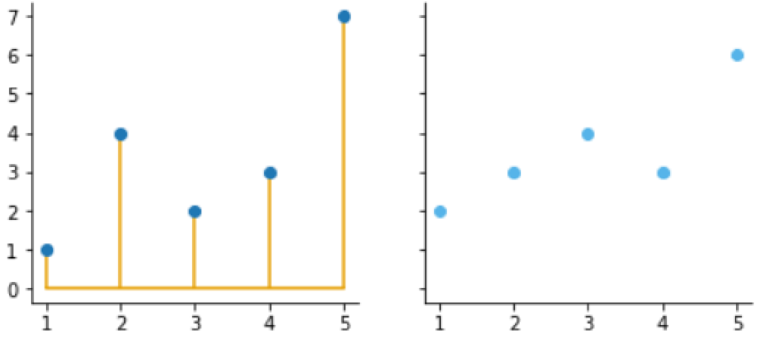
Bar plots



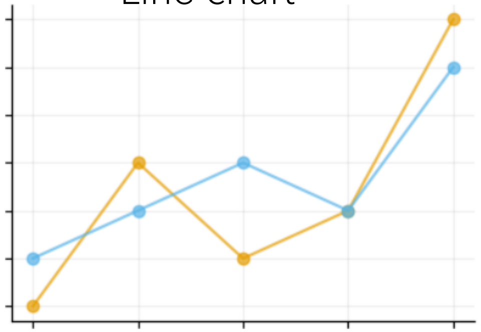
Heatmap



Dot plots



Line chart



Distributions

Distributions

- **Histograms** and **density** plots
- Careful! both require arbitrary parameter choices (e.g., # bins) and can be misleading

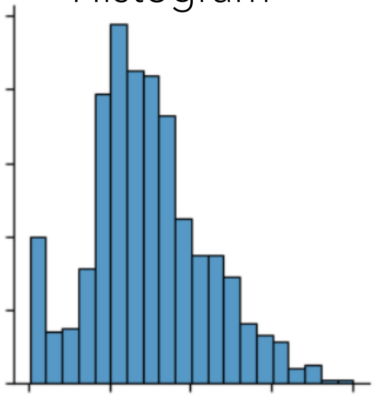
- **Cumulative densities** and **quantile-quantile** (q-q) plots always represent the data faithfully but can be more difficult to interpret.

Distributions

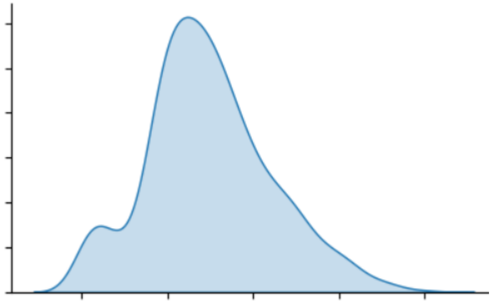
- **Histograms** and **density** plots
- Careful! both require arbitrary parameter choices (e.g., # bins) and can be misleading

- **Cumulative densities** and **quantile-quantile** (q-q) plots always represent the data faithfully but can be more difficult to interpret.

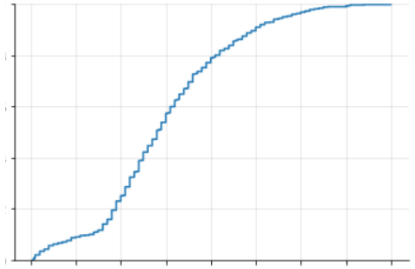
Histogram



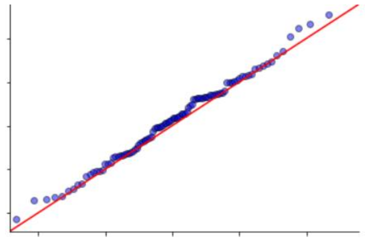
Density Plot



Cumulative Density



Quantile-Quantile Plot



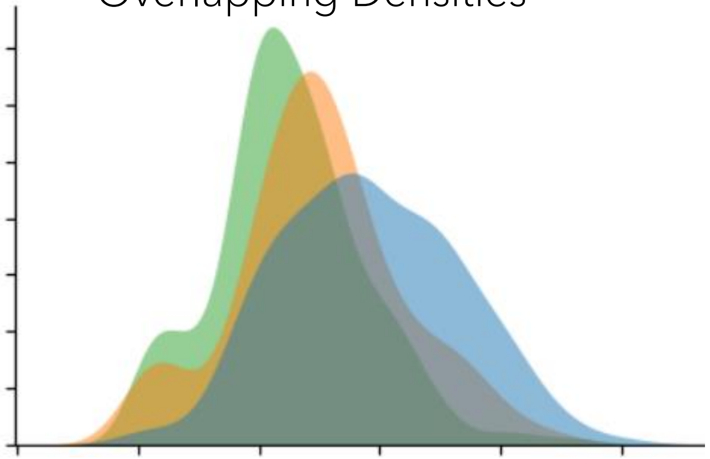
Multiple Distributions

- To visualize multiple distributions use tools such as boxplots, violin plots, overlapping densities

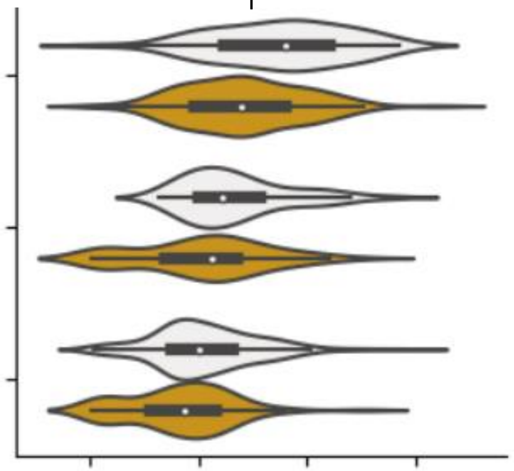
Multiple Distributions

- To visualize multiple distributions use tools such as boxplots, violins, overlapping densities

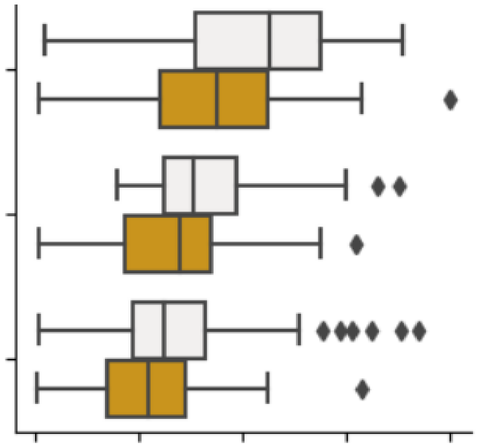
Overlapping Densities



Violin plots



Boxplots



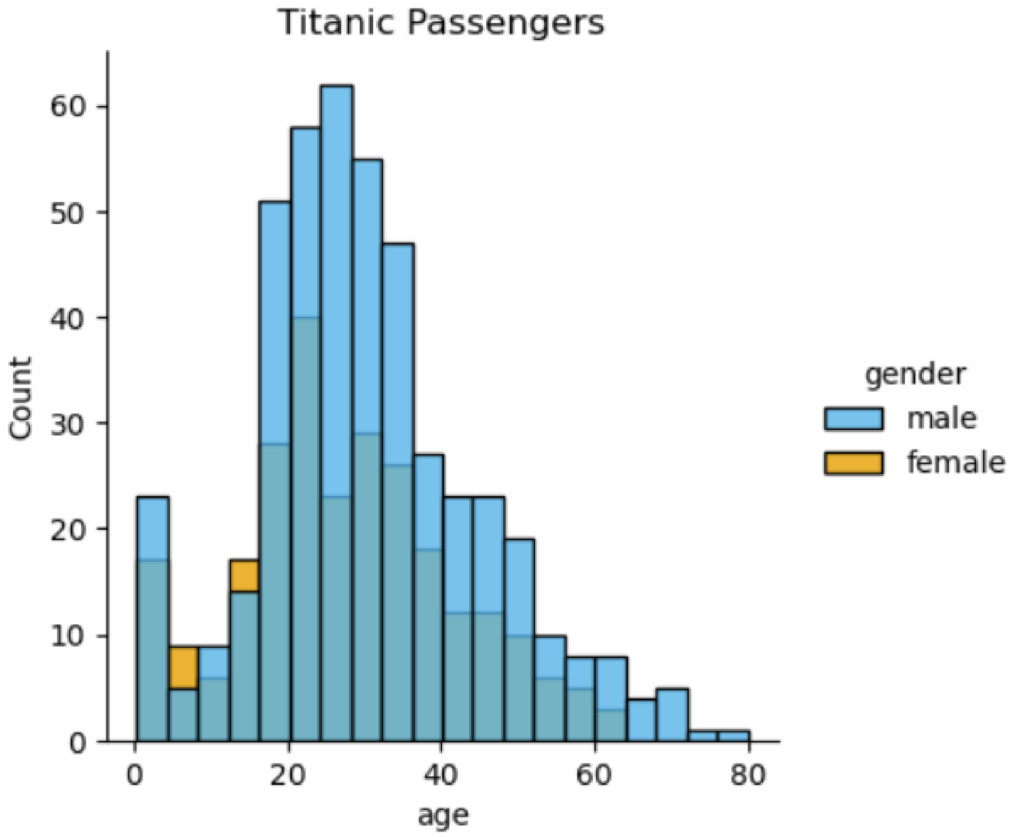
Multiple Distributions

- To visualize multiple distributions use tools such as boxplots, violins, overlapping densities

- You can also use **stacked or overlapping histograms**, but they may be difficult to interpret.

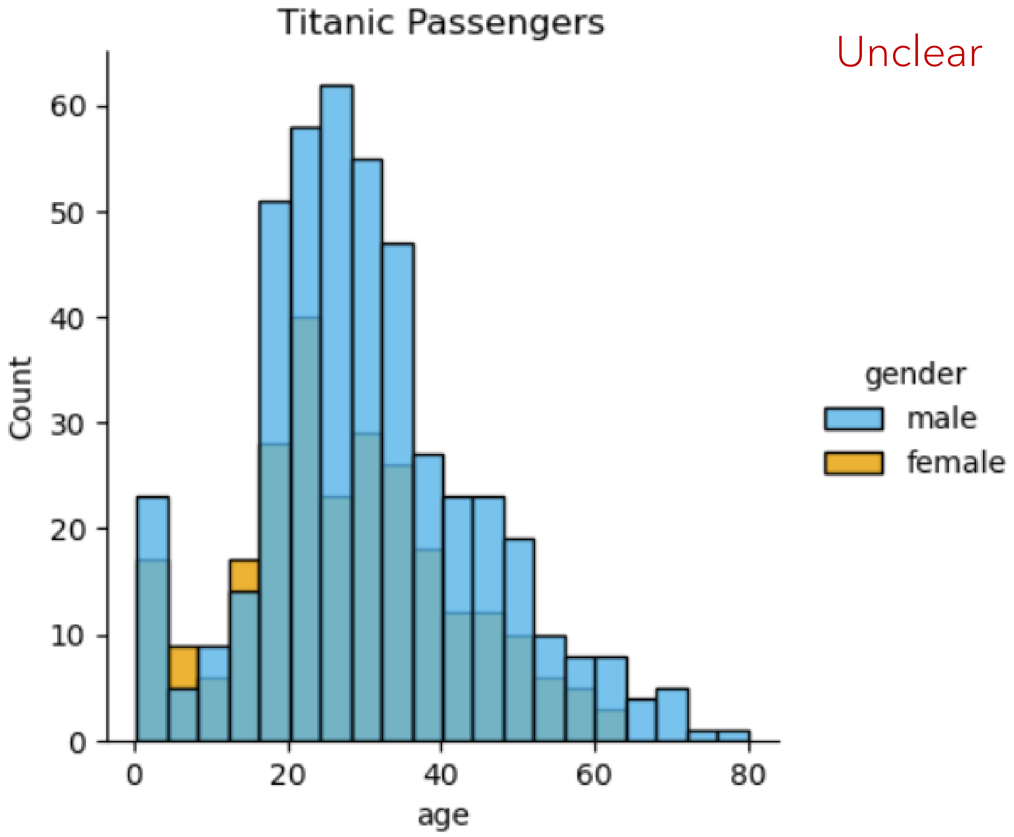
Multiple Distributions

- To visualize multiple distributions use tools such as boxplots, violins, overlapping densities
- You can also use **stacked or overlapping histograms**, but they may be difficult to interpret.



Multiple Distributions

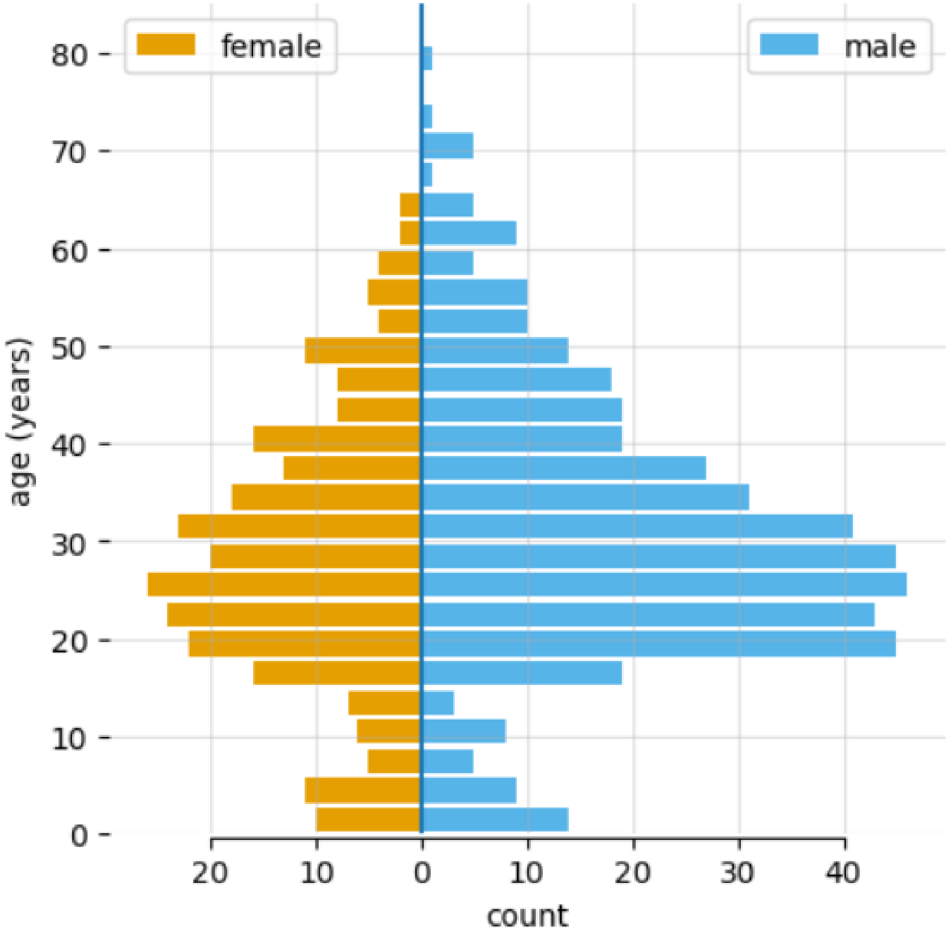
- To visualize multiple distributions use tools such as boxplots, violins, overlapping densities
- You can also use **stacked or overlapping histograms**, but they may be difficult to interpret.



Unclear

Multiple Distribution

- To visualize multiple distributions use tools such as boxplots, violins, overlapping densities
- You can also use **stacked or overlapping histograms**, but they may be difficult to interpret.



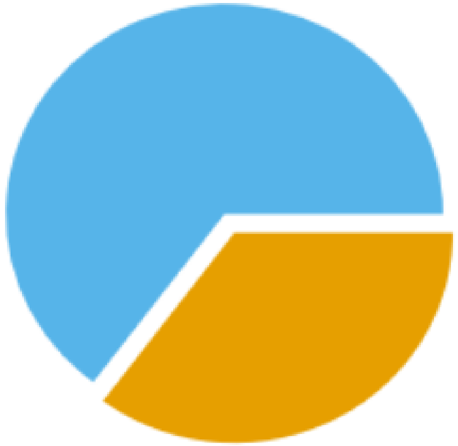
Better

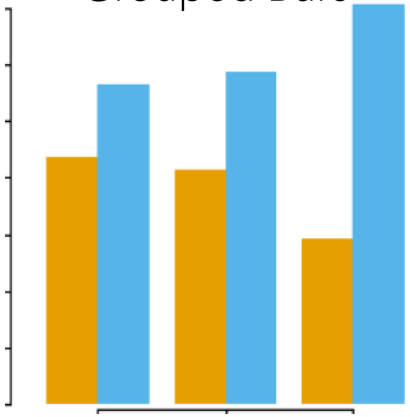
Proportions

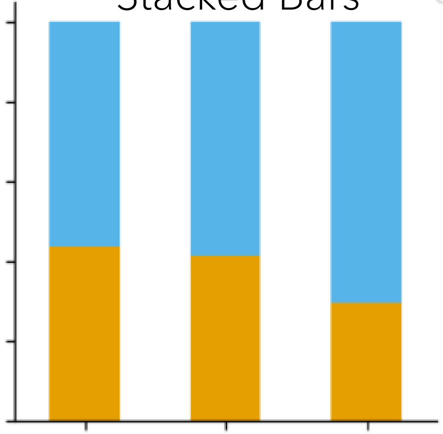
- Proportions can be visualized using pie charts and side-by-side bar or stacked bar plots.

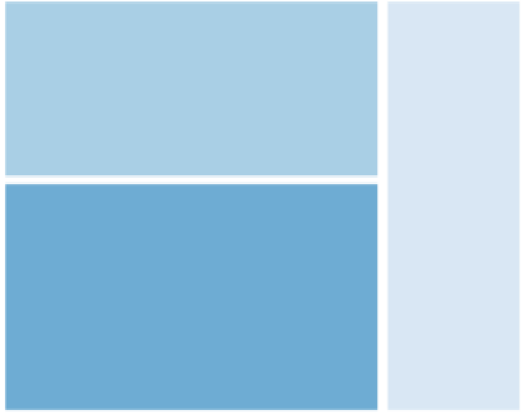
Proportions

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


Pie Chart


Grouped Bars


Stacked Bars


Treemap

Proportions

- When visualizing multiple sets of proportions across conditions, pie charts tend to be space-inefficient and often obscure relationships

Proportions

- When visualizing multiple sets of proportions across conditions, pie charts tend to be space-inefficient and often obscure relationships

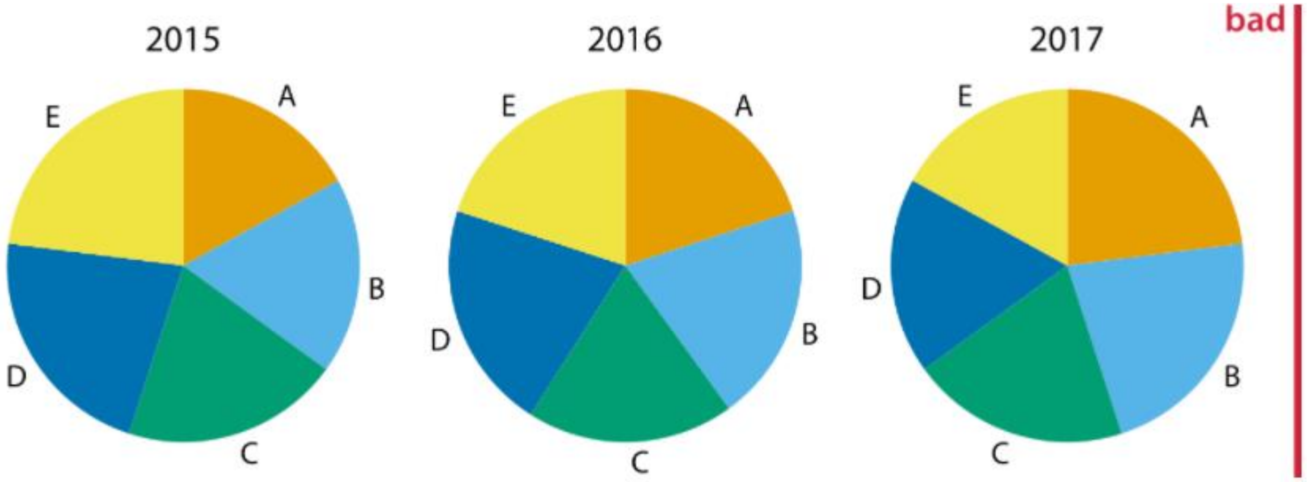


Figure 10.4: Market share of five hypothetical companies, A–E
Source: https://clauswilke.com/dataviz/visualizing-proportions.html

Proportions

- When visualizing multiple sets of proportions across conditions, pie charts tend to be space-inefficient and often obscure relationships. Consider bar plots.

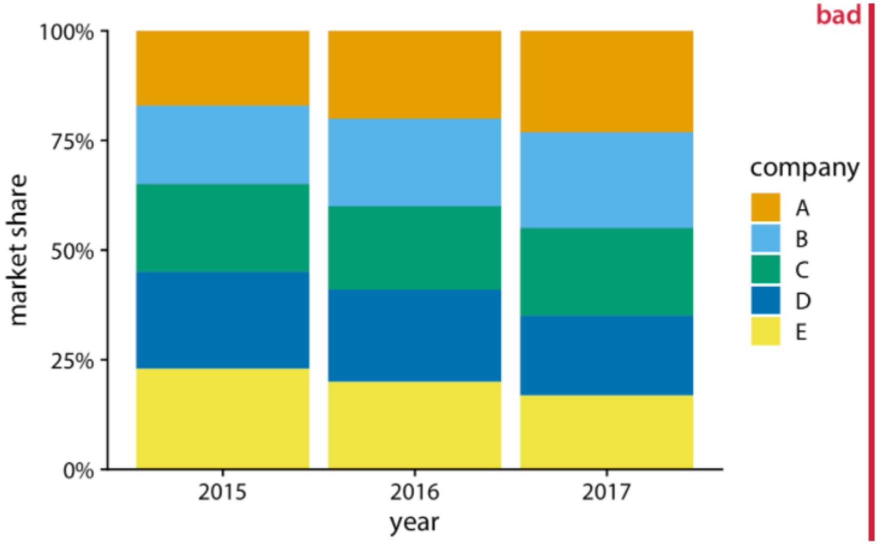


Figure 10.4: Market share of five hypothetical companies, A–E

Source: https://clauswilke.com/dataviz/visualizing-proportions.html

Proportions

- When visualizing multiple sets of proportions across conditions, pie charts tend to be space-inefficient and often obscure relationships. Consider bar plots.

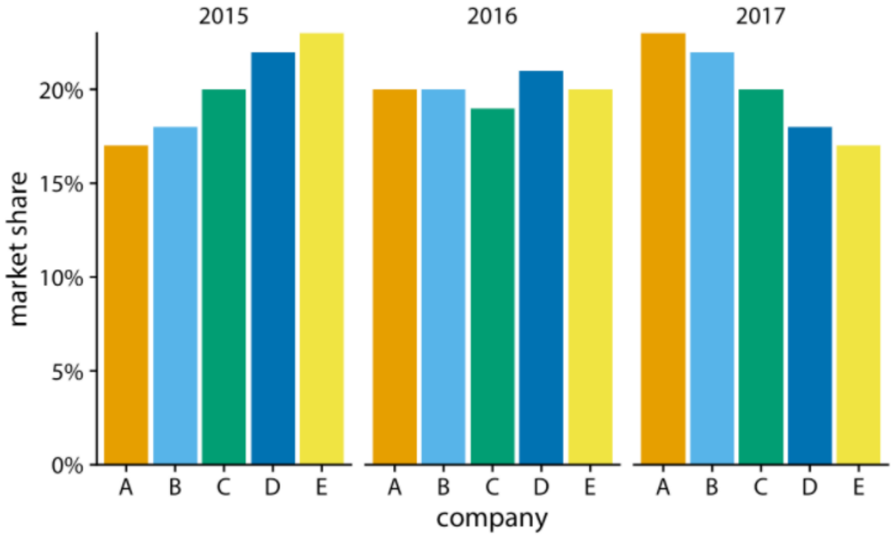


Figure 10.4: **Market share of five hypothetical companies, A–E**

Source: https://clauswilke.com/dataviz/visualizing-proportions.html

Proportions

- When proportions are specified according to multiple grouping variables, then **mosaic** plots, **treemap**s, or **parallel** sets are useful visualization approaches.

Proportions

- When proportions are specified according to multiple grouping variables, then **mosaic** plots, **treemaps**, or **parallel** sets are useful visualization approaches.
- How to represent proportions of bridge construction materials (steel, wood, iron) over specified categories (crafts, before 1870, modern, after 1940)?

Proportions

- When proportions are specified according to multiple grouping variables, then **mosaic** plots, **treemap**s, or **parallel** sets are useful visualization approaches.
- How to represent proportions of bridges construction materials (steel, wood, iron) over specified categories (crafts, before 1870, modern, after 1940)?

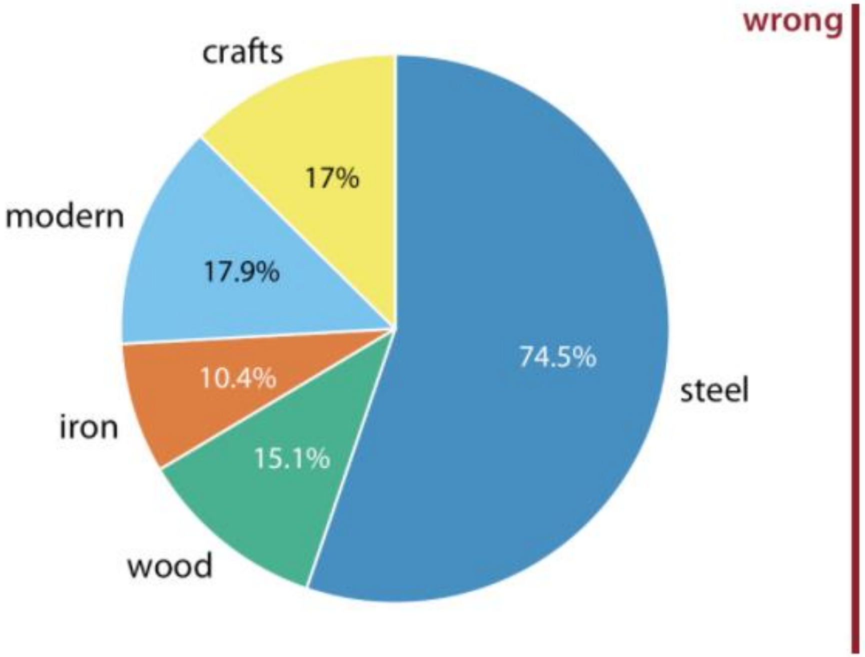


Figure 11.1: Breakdown of bridges in Pittsburgh **by construction material** (steel, wood, iron) and **by date of construction** (crafts, before 1870, and modern, after 1940). Source: https://clauswilke.com/dataviz/visualizing-proportions.html

Proportions

- When proportions are specified according to multiple grouping variables, then **mosaic** plots, **treemap**s, or **parallel** sets are useful visualization approaches.
- How to represent proportions of bridges construction materials (steel, wood, iron) over specified categories (crafts, before 1870, modern, after 1940)?

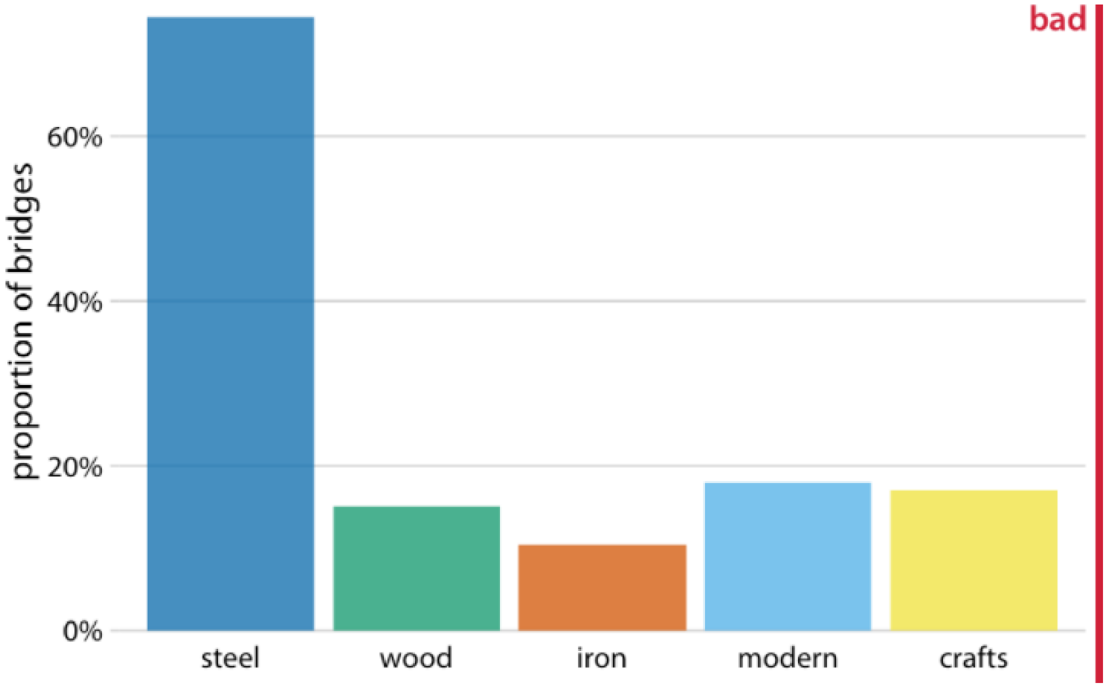


Figure 11.1: Breakdown of bridges in Pittsburgh **by construction material** (steel, wood, iron) and **by date of construction** (crafts, before 1870, and modern, after 1940). Source: https://clauswilke.com/dataviz/visualizing-proportions.html

Proportions

- When proportions are specified according to multiple grouping variables, then **mosaic** plots, **treemaps**, or **parallel** sets are useful visualization approaches.
- How to represent proportions of bridges construction materials (steel, wood, iron) over specified categories (crafts, before 1870, modern, after 1940)?

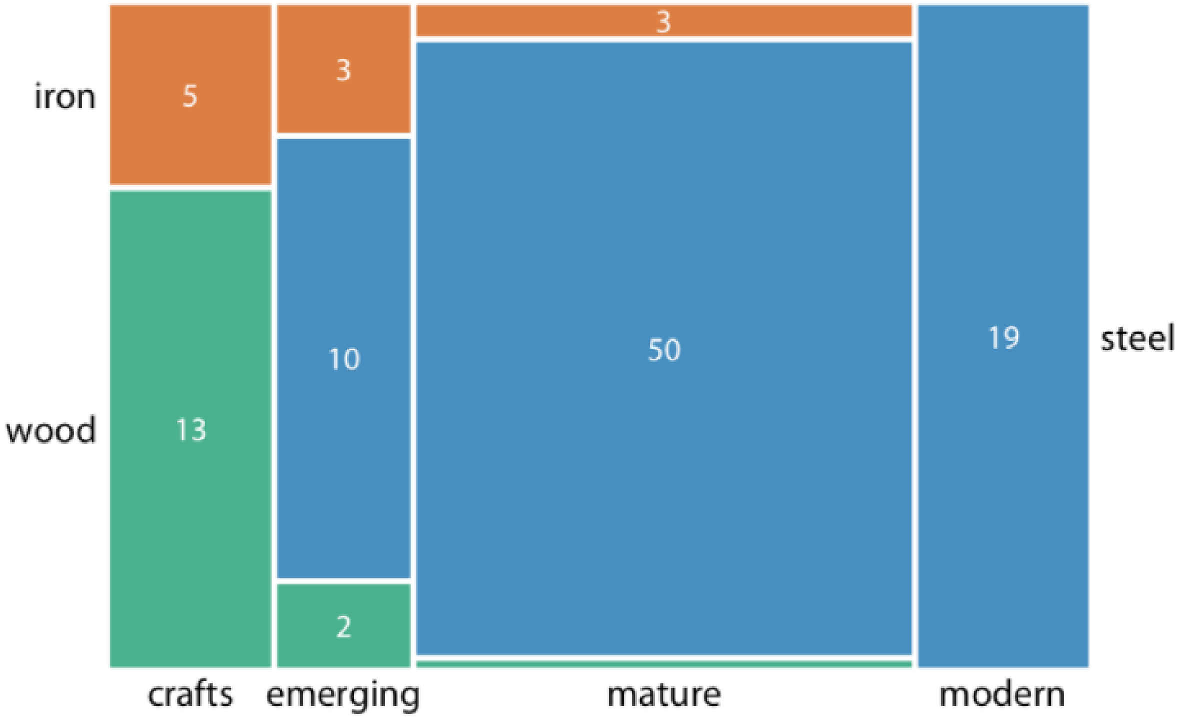


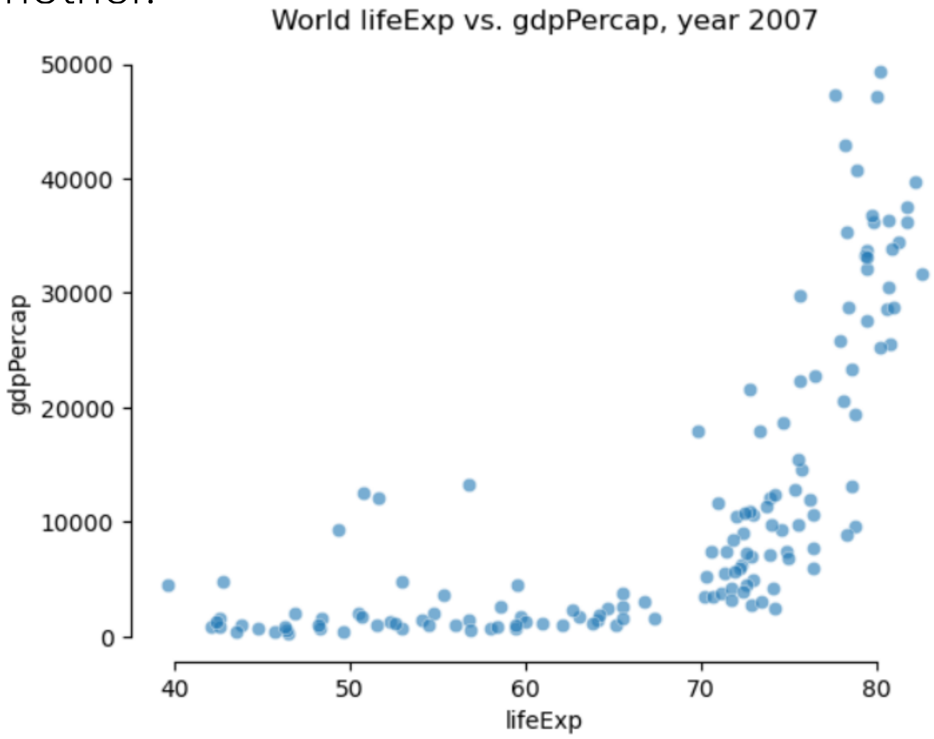
Figure 11.1: Breakdown of bridges in Pittsburgh **by construction material** (steel, wood, iron) and **by date of construction** (crafts, before 1870, and modern, after 1940). Source: https://clauswilke.com/dataviz/visualizing-proportions.html

x–y relationships

- Scatterplots represent the archetypical visualization when we want to show one quantitative variable relative to another.

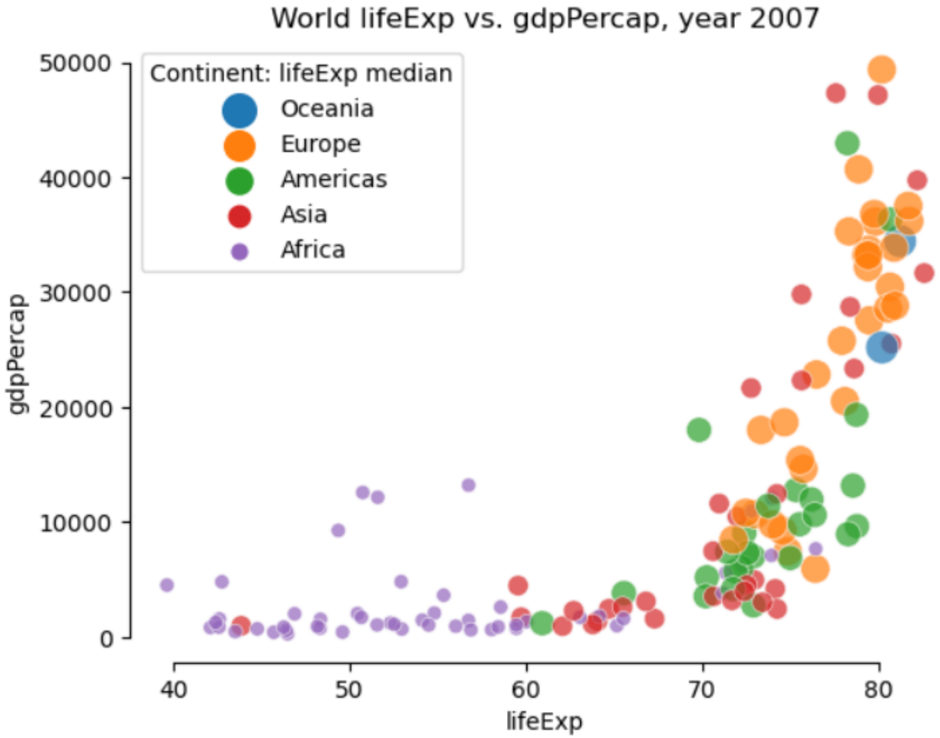
x–y relationships

- **Scatterplots** represent the archetypical visualization when we want to show one quantitative variable relative to another.



x–y relationships

- If we have three quantitative variables, we can map one onto the dot size, creating a variant of the scatterplot called **bubble chart.**



x–y relationships

- For paired data, where the variables along the x and the y axes are measured in the same units, it is generally helpful to add a line indicating x = y (**paired scatterplot**)

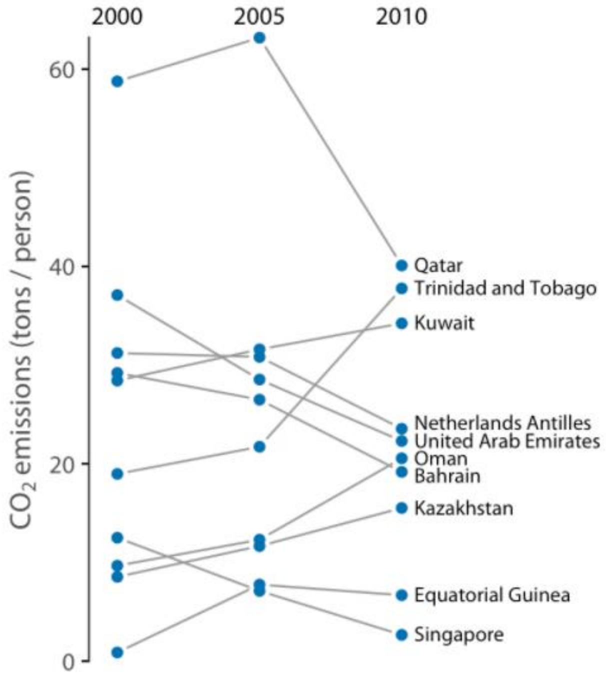


Figure 12.13: **CO2 emissions per person in 2000, 2005, and 2010,** for the ten countries with the largest difference between the years 2000 and 2010 Source: https://clauswilke.com/dataviz/visualizing-associations.html

The larger picture

Telling a story and making a point

- Most data visualization is done for the purpose of communication.

- To communicate your insight successfully, you must present the audience with a clear and exciting story.

- Each figure you make should be part of your overall story.

- It is possible to tell the whole story with one figure

Make a figure for the generals

- Never assume your audience can rapidly process complex visual displays.
- Refrain from attempting to show too much information in one figure
- Keep your figures simple and avoid confusing labels and overly technical terms
- After all, the generals are simply very busy!

Make a figure for the generals

- Avoid complex figures

Make a figure for the generals

- Too complex

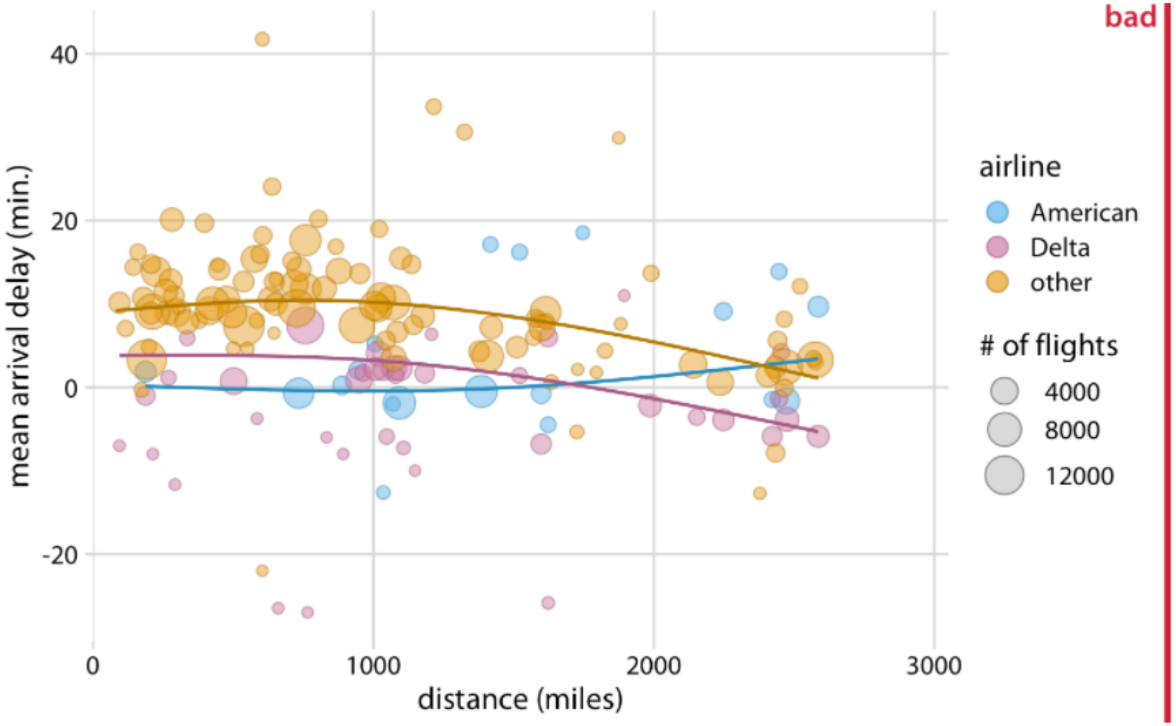


Figure 29.3: Mean arrival delay versus distance from New York City. Each point represents one destination, and the size of each point represents the number of flights from one of the three major New York City airports (Newark, JFK, or LaGuardia) to that destination in 2013. Source: https://clauswilke.com/dataviz/telling-a-story.html#what-is-a-story

Build up towards complex figures

- If you have a lot of information to share, start by showing a simplified version of the story.

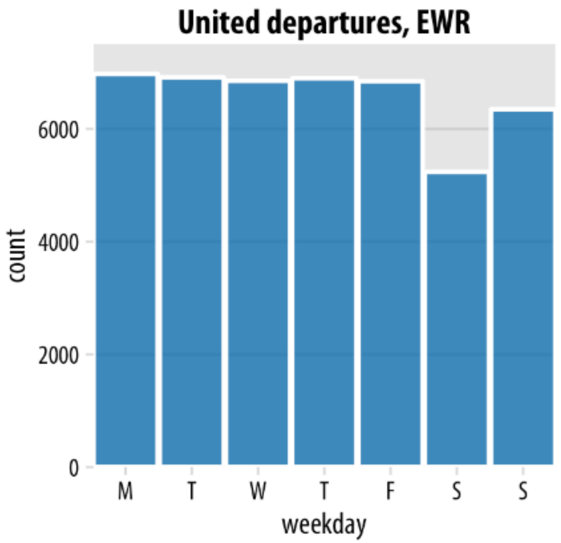
- For example, before showing a figure with multiple subplots, first show one plot

- Say you are trying to tell a story about the variation in the departures of multiples airlines by the day of the week.

- How to proceed?

Build up towards complex figures

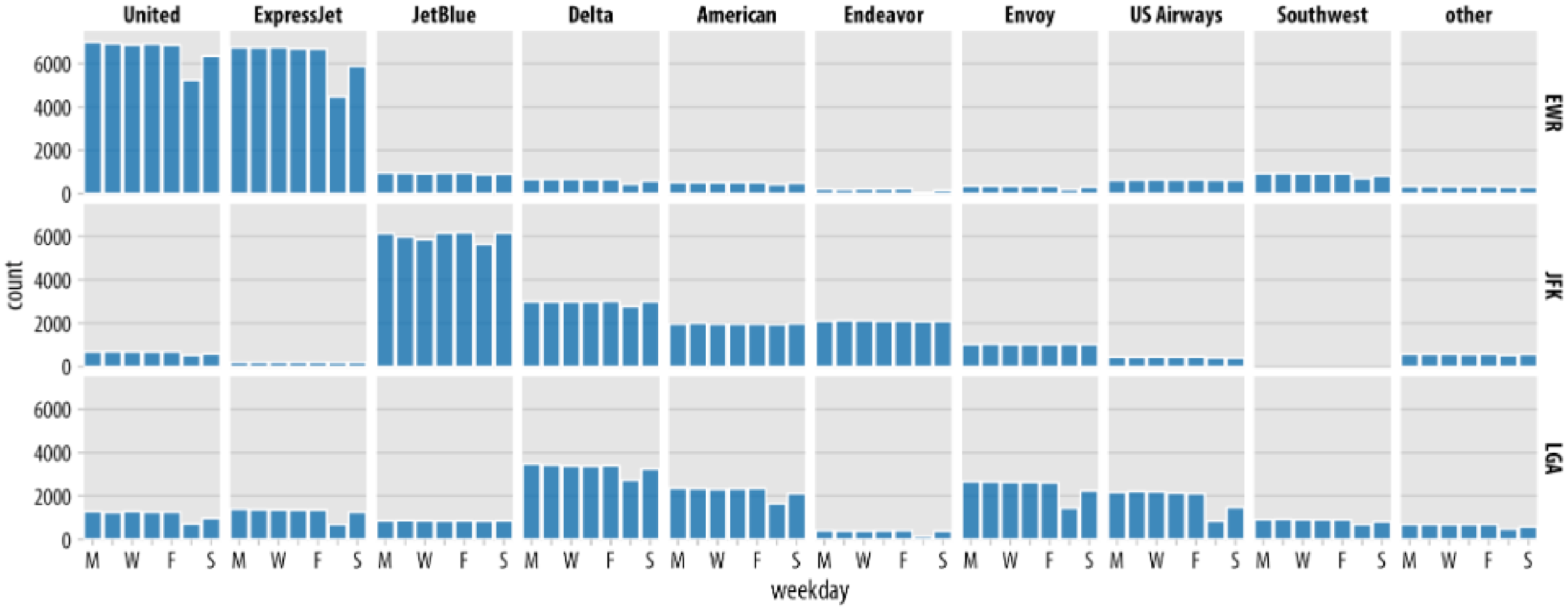
- You can start by showing a figure of the airline that stands our the most



United Airlines departures out of Newark Airport (EWR) in 2013
(Source: Figure 29.6 https://clauswilke.com/dataviz/)

Build up towards complex figures

- Next, you can present a full grid of subplots



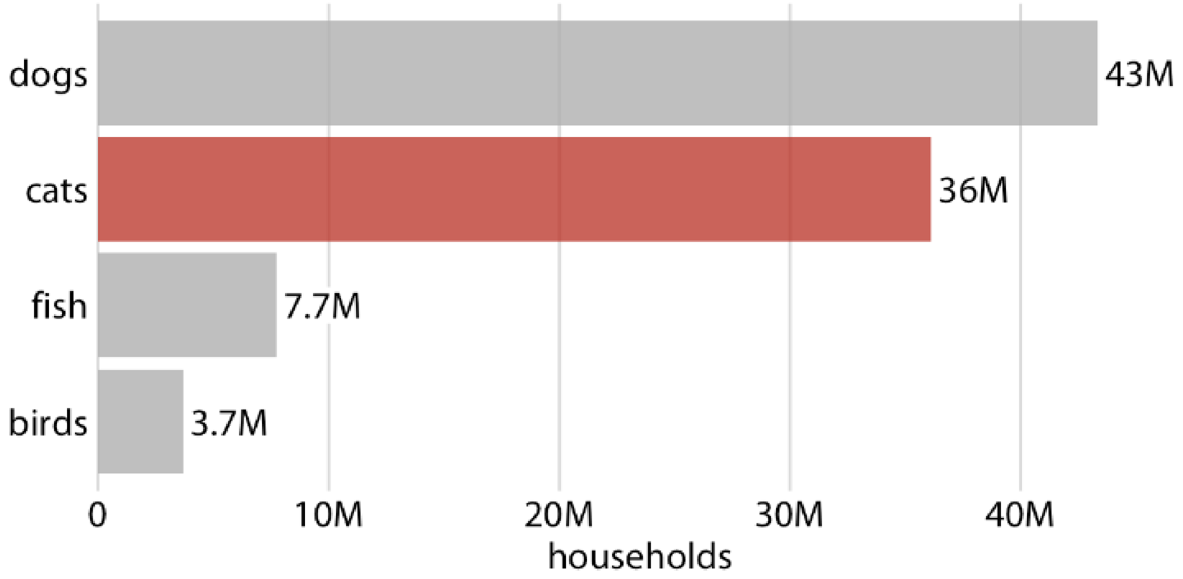
Departures out of airports in the New York city area in 2013
(Source: Figure 29.7 https://clauswilke.com/dataviz/)

Make your figures memorable

- If you have a good story to tell, make it memorable.
- Again, simple figures **without complex information** are recommended
- For example, **whenever applicable, use bar plots** since they are easily interpreted.
- Simple **does not mean generic**

Make your figures memorable

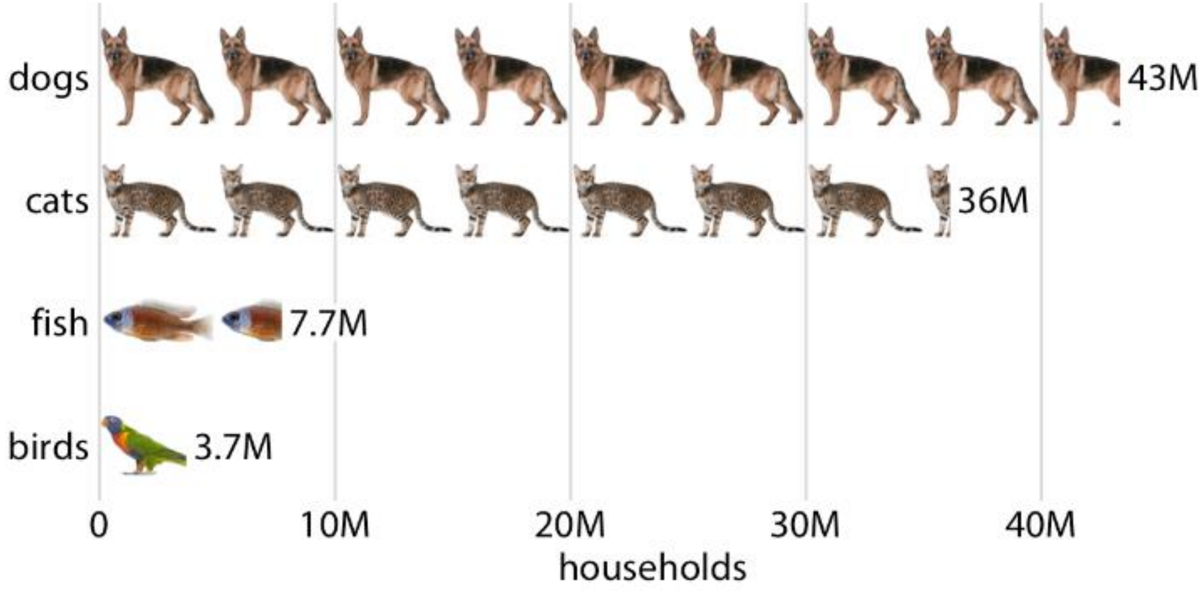
- A story about cats



Number of households having one or more of the most popular pets in 2012
(Source: Figure 29.8 https://clauswilke.com/dataviz/)

Make your figures memorable

- A story about cats



Number of households having one or more of the most popular pets in 2012
(Source: Figure 29.9 https://clauswilke.com/dataviz/)

Be consistent but don’t be repetitive

- If figures are part of the same story, they should look like they belong in the same story.

- This does not mean using the same type of chart throughout--your story may end up being confusing and boring.

Be consistent but don’t be repetitive

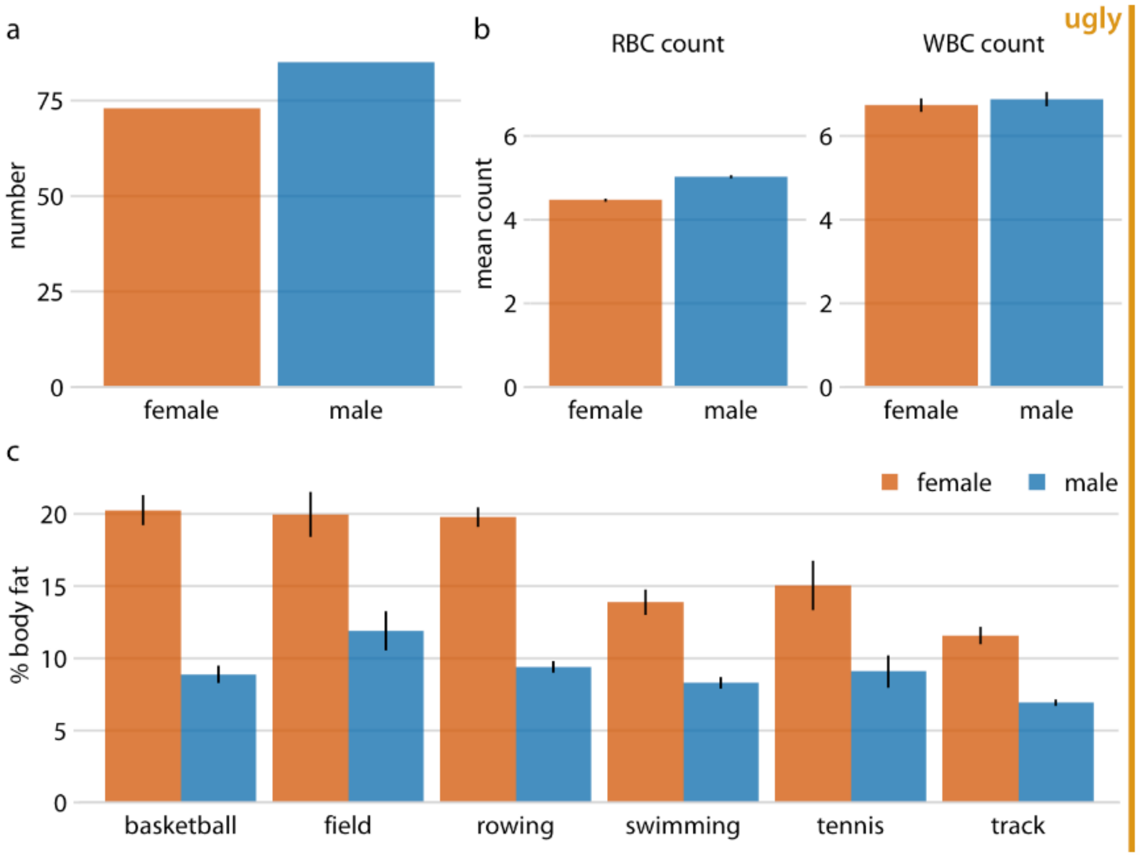
- If figures are part of the same story, they should look like they belong in the same story.

- This does not mean using the same type of chart throughout--your story may end up being confusing and boring.

Be consistent but don’t be repetitive

- If figures are part of the same story, they should look like they belong in the same story.

- This does not mean using the same type of chart throughout--your story may end up being **confusing** and boring.



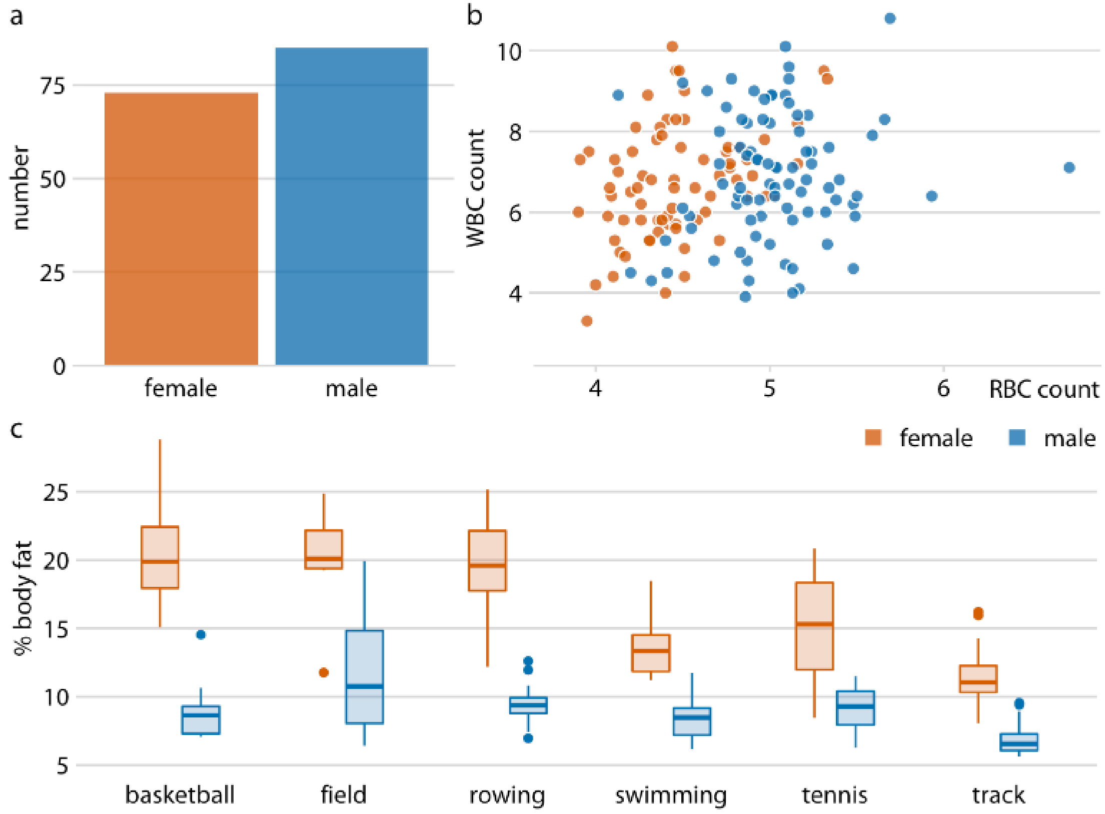
Physiology and body-composition of male and female athletes

(Source: Figure 29.10 https://clauswilke.com/dataviz/)

Be consistent but don’t be repetitive

- If figures are part of the same story, they should look like they belong in the same story.

- Better!



Physiology and body-composition of male and female athletes

(Source: Figure 21.8 https://clauswilke.com/dataviz/)

Be consistent but don’t be repetitive

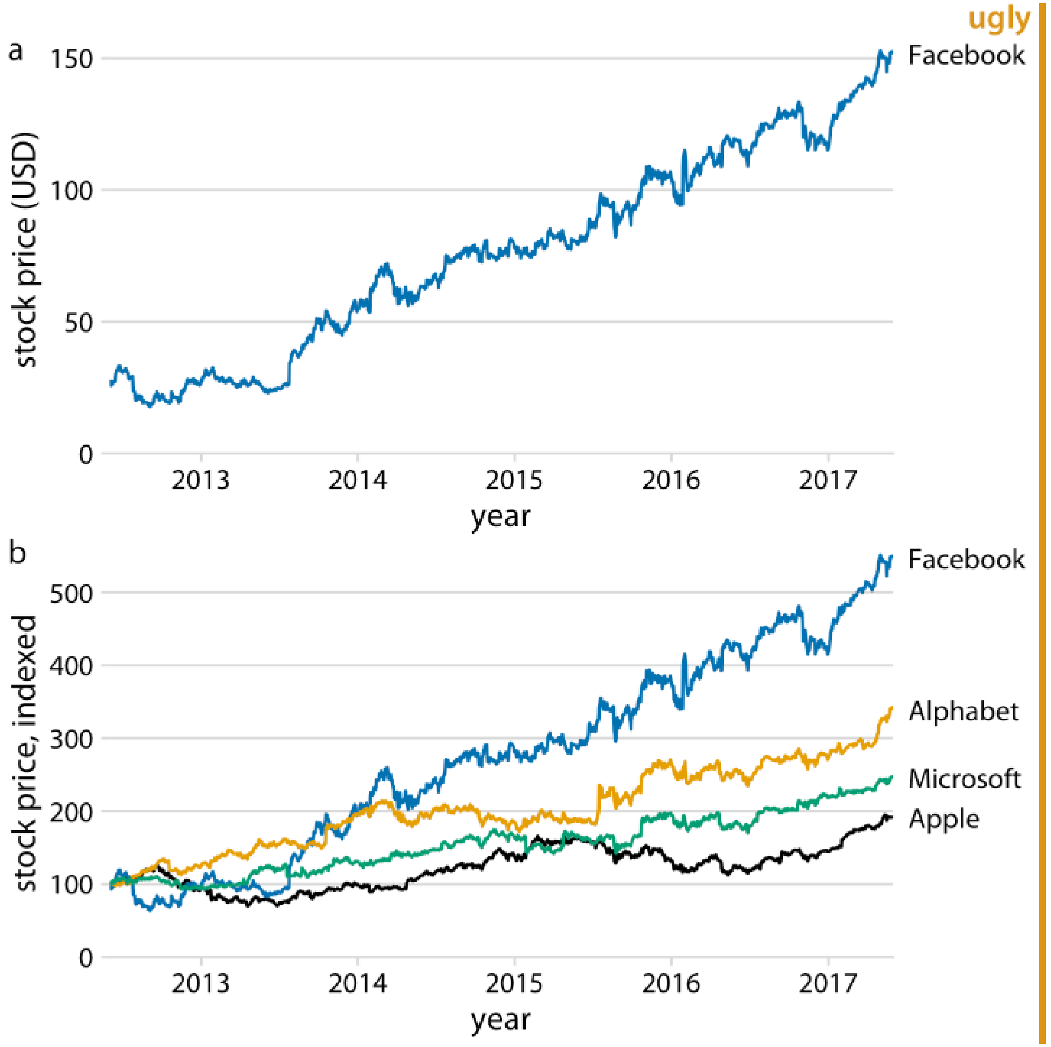
- If you **used a line chart before**, consider using other easy to interpret plots such as boxplots, line charts, or scatter plots.

- Stay consistent with the **color** and only highlight what is important to your story

Be consistent but don’t be repetitive

- If you used a line chart before, consider using other easy to interpret plots such as boxplots, line charts, or scatter plots.

- Stay consistent with the color and only highlight what is important to your story



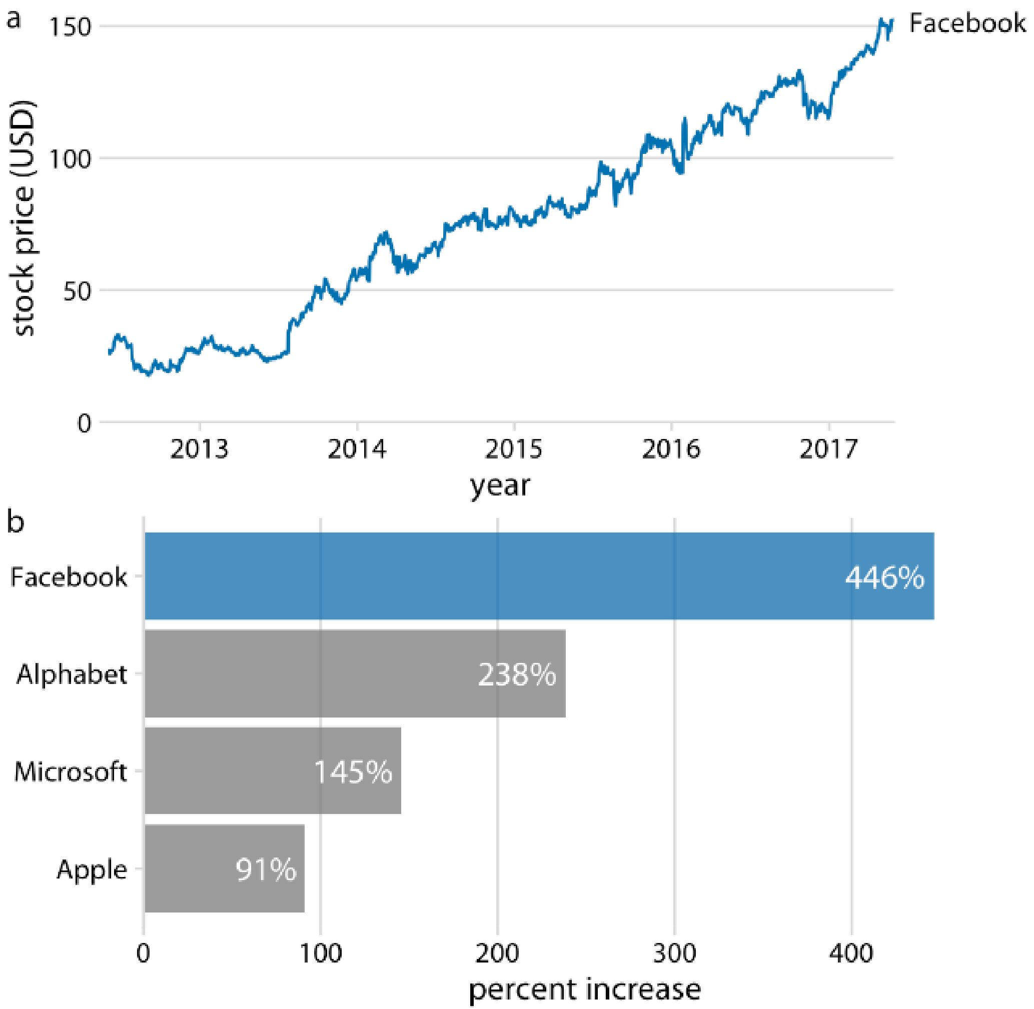
Growth of Facebook stock price over a five-year interval in comparison with other tech stocks (Source: Figure 29.11 https://clauswilke.com/dataviz/)

Be consistent but don’t be repetitive

- If you used a bar chart before, consider using other easy to interpret plots such as boxplots, line charts, or scatter plots.

- Stay consistent with the color and only highlight what is important to your story

- Better!



Growth of Facebook stock price over a five-year interval and comparison with other tech stocks (Source: Figure 29.11 https://clauswilke.com/dataviz/)

Common pitfalls of color use

Color scales

- Why do we use colors?

We use color to distinguish groups of data from each other

We use color to represent data values

We use color to highlight specific elements

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

Common pitfalls of color use: lessons

1. Avoid encoding too much or irrelevant information

Common pitfalls of color use

- Poor color choices can ruin an otherwise excellent visualization

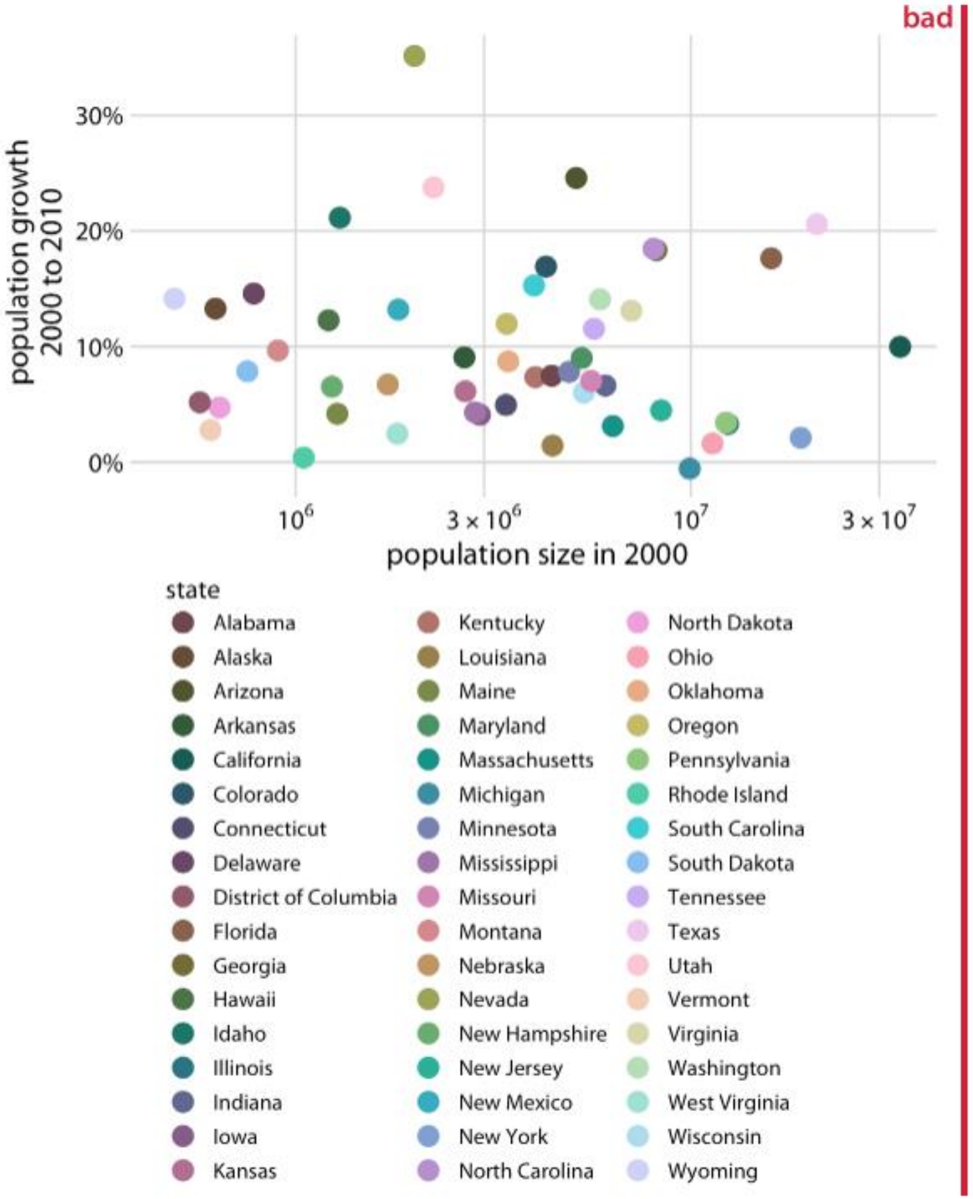
- Encoding too many different items in different colors is bad

Common pitfalls of color use

- Poor color choices can ruin an otherwise excellent visualization

- Encoding too many different items in different colors is bad

- Which State is which?



Population growth from 2000 to 2010 versus population size in 2000,for all 50 U.S. states and the District of Columbia (Source: Figure 19.1 https://clauswilke.com/dataviz/)

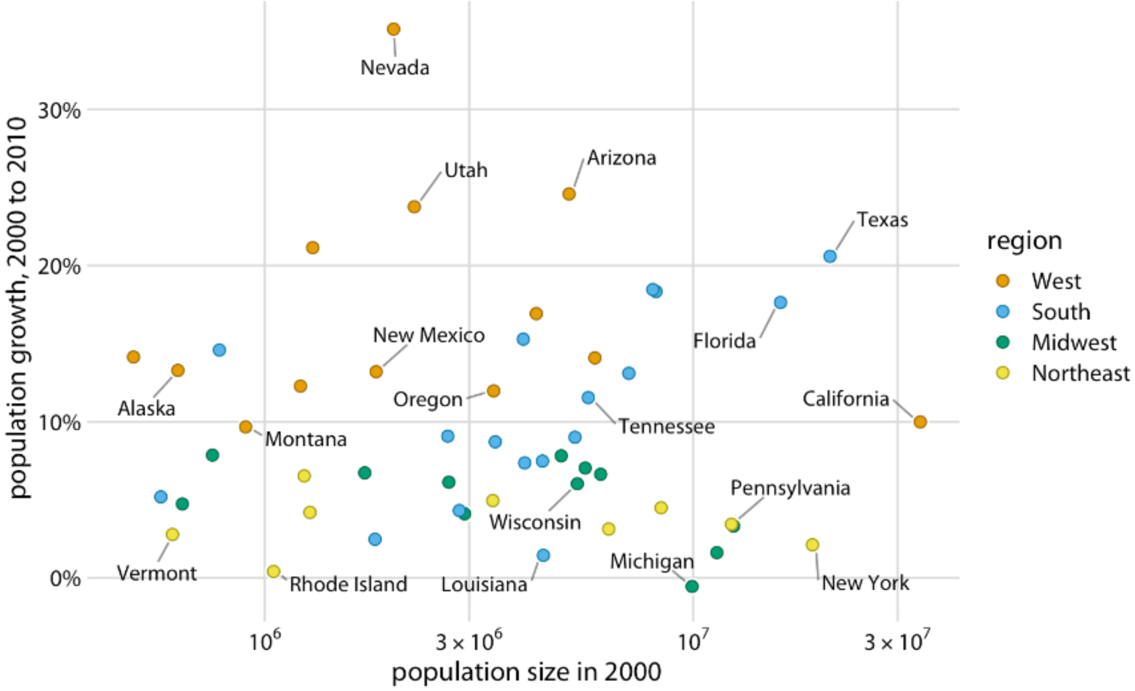
Common pitfalls of color use

- Rule of thumb: Qualitative color scales work best when there are three to five different categories that need to be colored

Common pitfalls of color use

- Rule of thumb: Qualitative color scales work best when there are three to five different categories that need to be colored

- One way to fix this figure is to label a re presentative set of geographical US regions

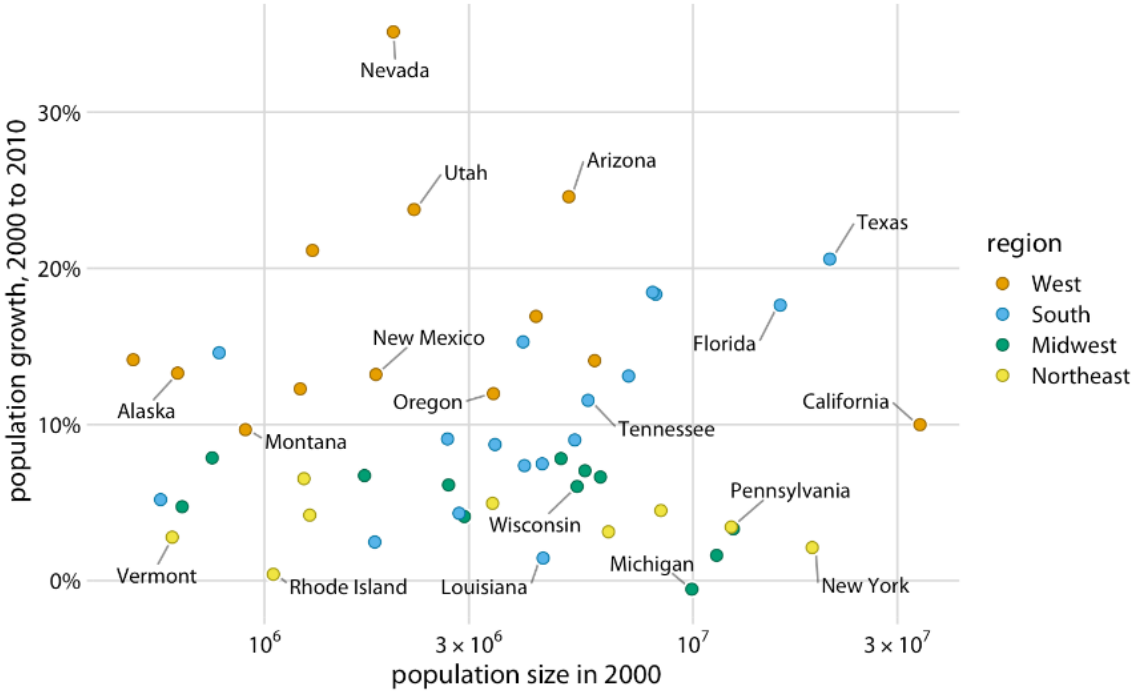


Population growth from 2000 to 2010 versus population size in 2000,for all 50 U.S. states and the District of Columbia (Source: Figure 19.1 https://clauswilke.com/dataviz/)

Common pitfalls of color use

- Rule of thumb: Qualitative color scales work best when there are three to five different categories that need to be colored

- One way to fix this figure is to label a re presentative set of geographical US regions
- Use direct labeling instead of colors when you need to distinguish between more than about eight categorical items.



Population growth from 2000 to 2010 versus population size in 2000,for all 50 U.S. states and the District of Columbia (Source: Figure 19.1 https://clauswilke.com/dataviz/)

Common pitfalls of color use:
lessons

1. Avoid encoding too much or irrelevant information
2. Avoid coloring for the sake of coloring

Common pitfalls of color use

- Avoid large filled areas of **overly saturated colors** as they may confuse the audience

- Think of the purpose of the coloring

  Distinguish items or represent data value

  Highlight or draw attention to

  To tell a story!

Common pitfalls of color use

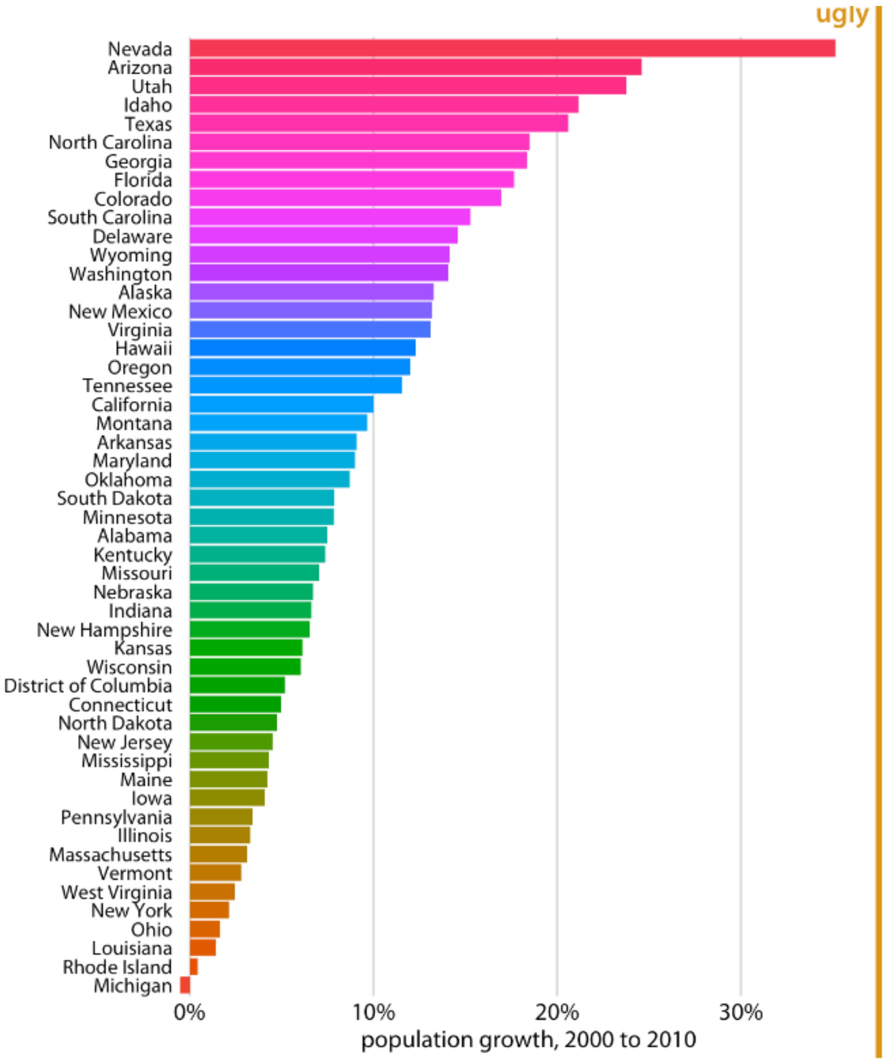
- Avoid coloring for the sake of coloring
- Avoid large filled areas of overly saturated colors as they may confuse the audience
- Think of the purpose of the coloring

  Distinguish items

  Highlight or draw attention to

  To tell a story

  The rainbow coloring of states serves no purpose and is distracting



Population growth from 2000 to 2010 versus population size in 2000,for all 50 U.S. states and the District of Columbia (Source: Figure 19.3 https://clauswilke.com/dataviz/)

Common pitfalls of color use

- Avoid coloring for the sake of coloring

- Avoid large filled areas of overly saturated colors as they may confuse the audience

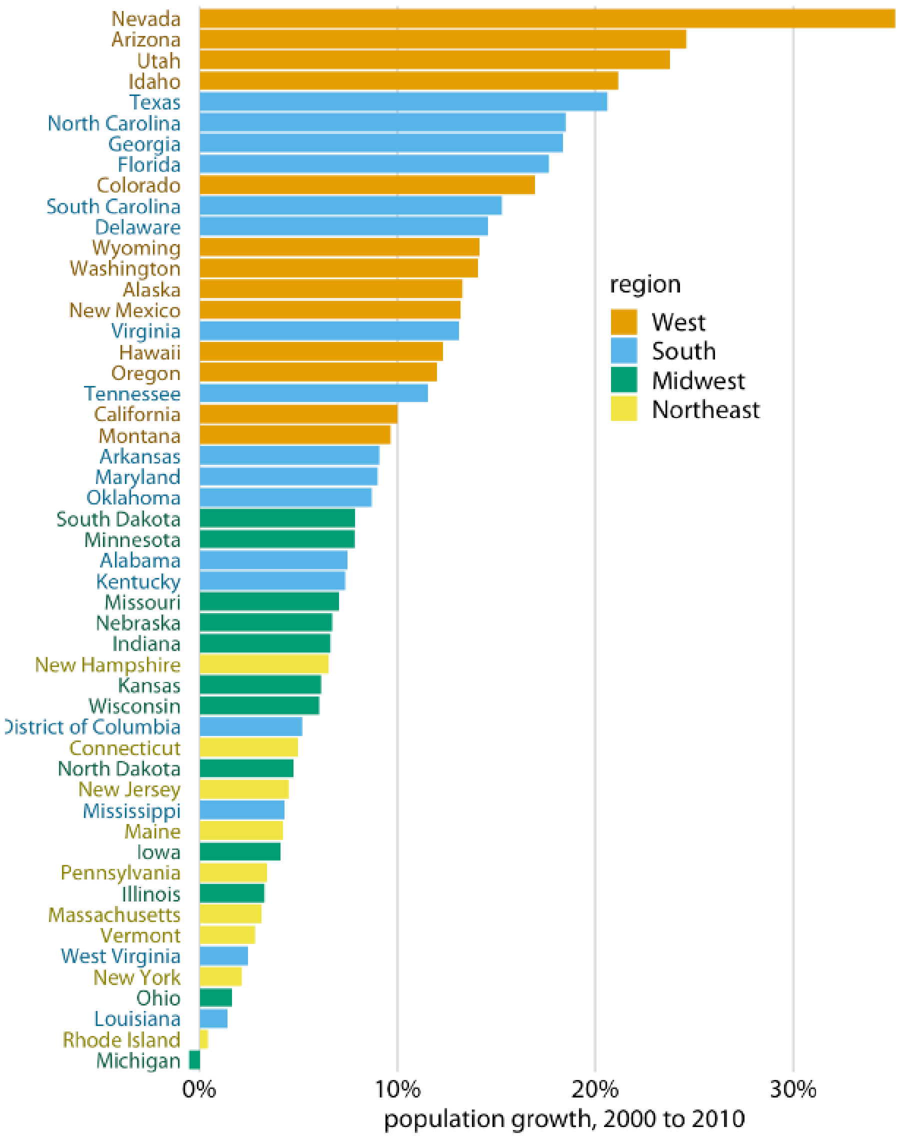
- Think of the purpose of the coloring

  Distinguish items

  Highlight or draw attention to

  To tell a story

  Make the color meaningful



Population growth from 2000 to 2010 versus population size in 2000,for all 50 U.S. states and the District of Columbia (Source: Figure 4.2 https://clauswilke.com/dataviz/)

Common pitfalls of color use

- What if you wanted to highlight certain states?

Common pitfalls of color use

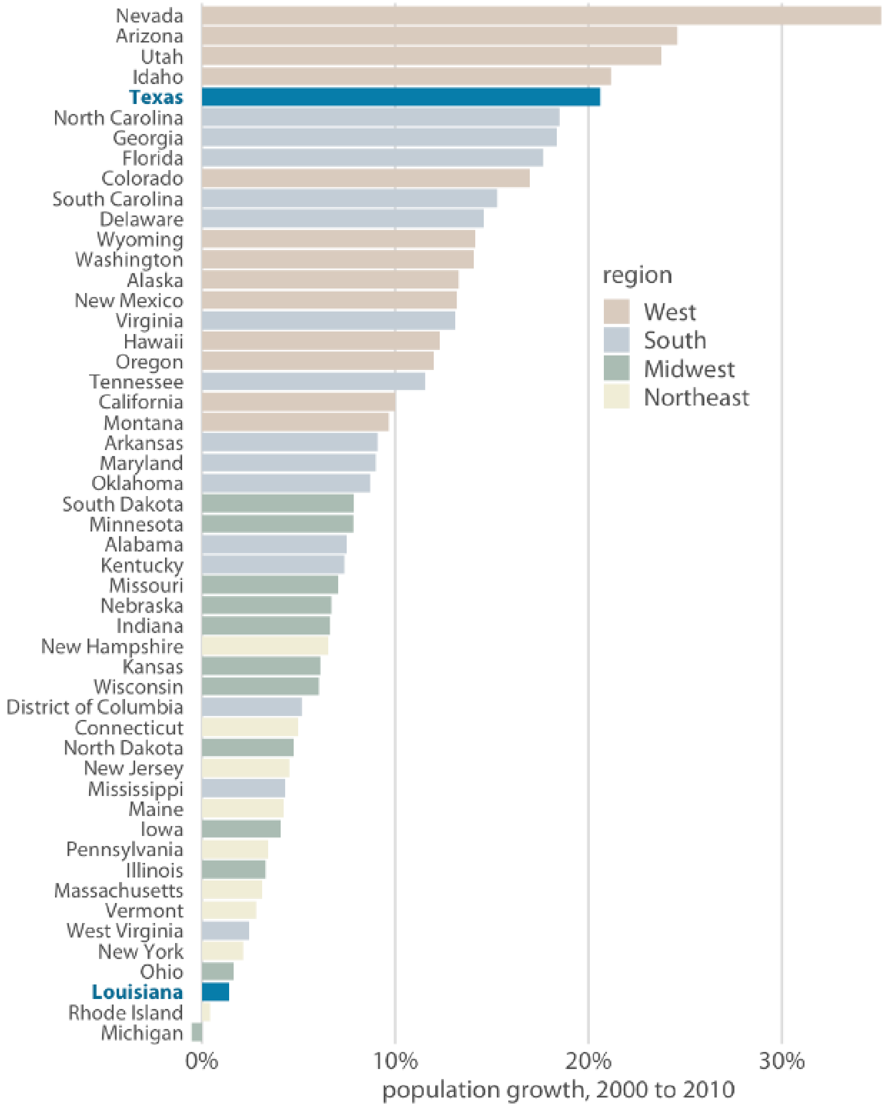
- What if you wanted to highlight certain states?
- Use an accent color scale

Common pitfalls of color use

- What if you wanted to highlight certain states?
- Use an accent color scale
- Utilizes a set of subdued colors and a matching set of stronger, darker, and/or more saturated colors

Common pitfalls of color use

- What if you wanted to highlight certain states?
- Use an accent color scale
- Utilizes a set of subdued colors and a matching set of stronger, darker, and/or more saturated colors



Population growth from 2000 to 2010 versus population size in 2000,for all 50 U.S. states and the District of Columbia (Source: Figure 4.2 https://clauswilke.com/dataviz/)

Common pitfalls of color use:
lessons

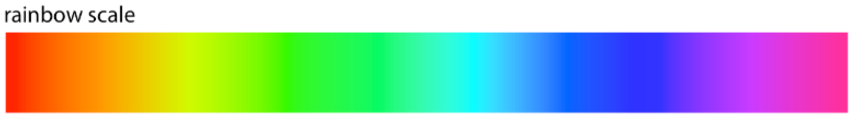
1. Avoid encoding too much or irrelevant information
2. Avoid coloring for the sake of coloring
3. Using monotonic color scales to encode data values

Common pitfalls of color use

- The colors need to clearly indicate which data **values are larger or smaller** than which other ones
- Differences between colors need to visualize the corresponding differences between data values

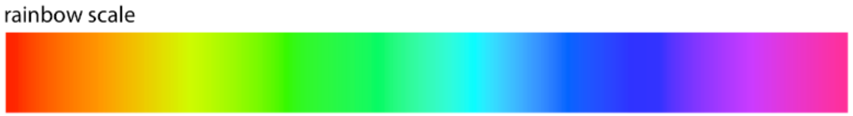
Common pitfalls of color use

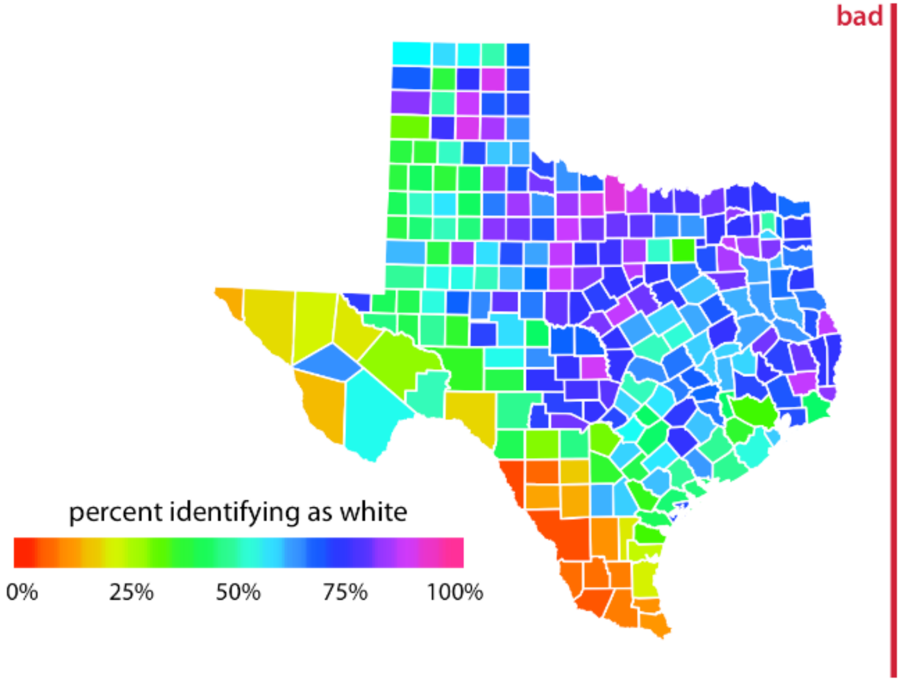
- The colors need to clearly indicate which data values are larger or smaller than which other ones
- Differences between colors need to visualize the corresponding differences between data values
- Avoid the non-monotonic scales such as the rainbow scale



Common pitfalls of color use

- The colors need to clearly indicate which data values are larger or smaller than which other ones
- Differences between colors need to visualize the corresponding differences between data values
- Avoid the non-monotonic scales such as the rainbow scale

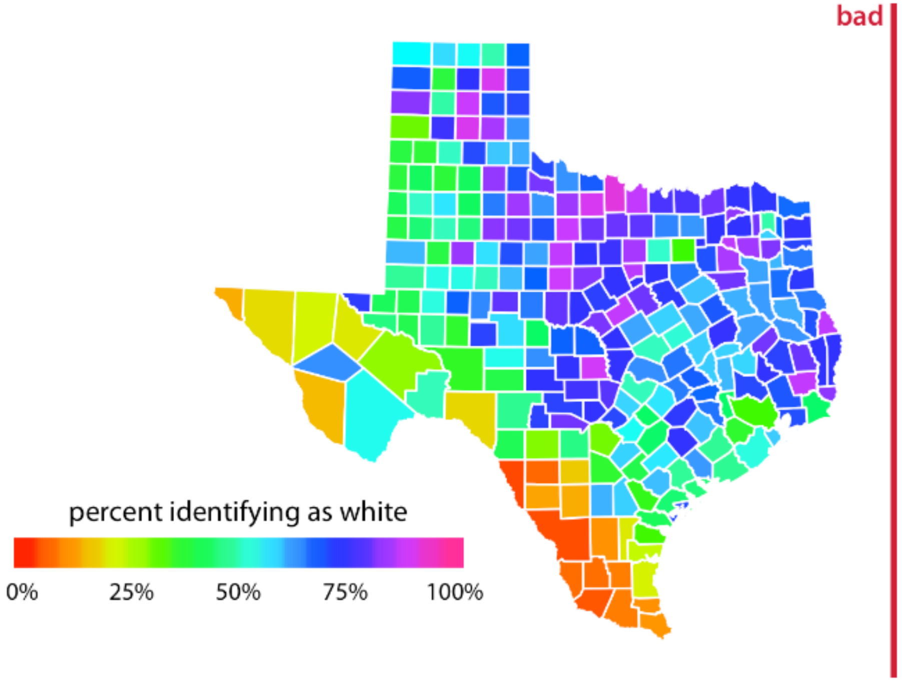




Percentage of people identifying as white in Texas counties (Source: Figure 19.5 https://clauswilke.com/dataviz/)

Common pitfalls of color use

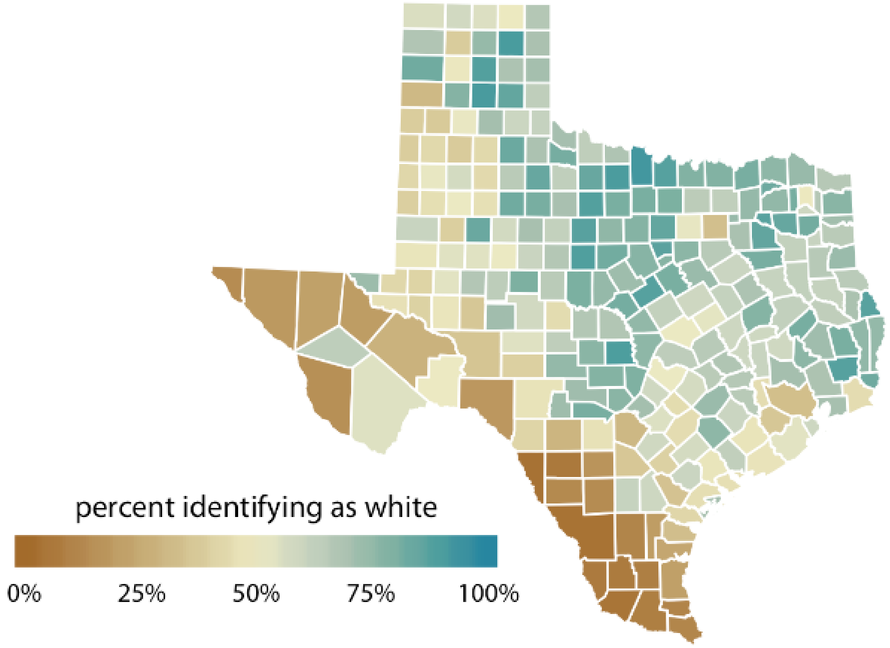
- The colors need to clearly indicate which data values are larger or smaller than which other ones
- Differences between colors need to visualize the corresponding differences between data values
- Avoid the non-monotonic scales such as the rainbow scale
- Too saturated for the eye
- Not an appropriate scale to visualize continuous data values



Percentage of people identifying as white in Texas counties (Source: Figure 19.5 https://clauswilke.com/dataviz/)

Common pitfalls of color use

- The colors need to clearly indicate which data values are larger or smaller than which other ones
- Differences between colors need to visualize the corresponding differences between data values
- Avoid the non-monotonic scales such as the rainbow scale
- Instead of the ~~rainbow scale~~, try a diverging color scale to represent data values about the identify of the population in Texas.

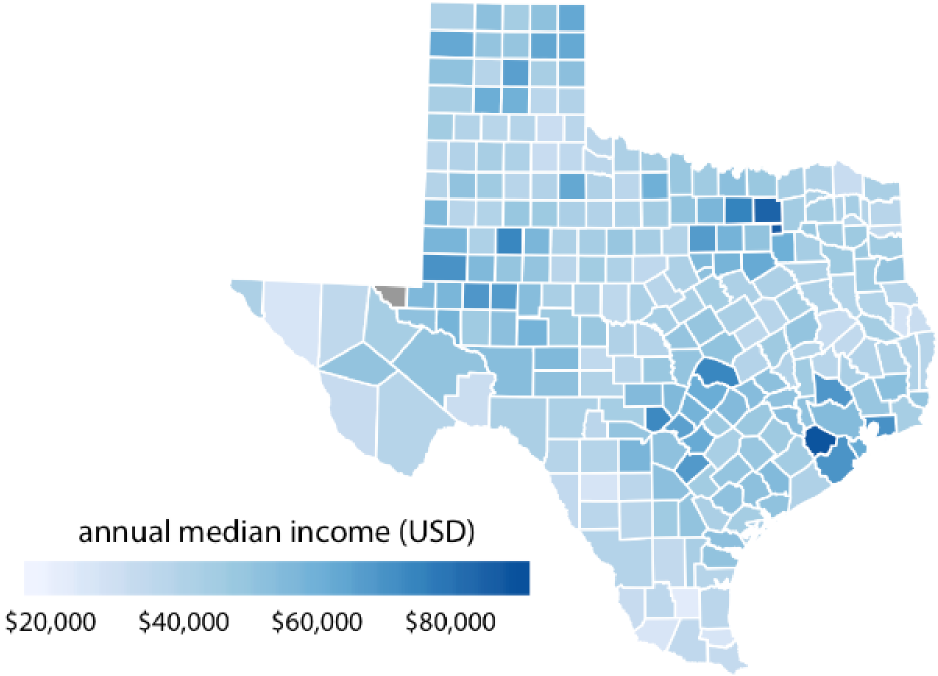


Percentage of people identifying as white in Texas counties

(Source: Figure4.6 https://clauswilke.com/dataviz/)

Common pitfalls of color use

- The colors need to clearly indicate which data values are larger or smaller than which other ones
- Differences between colors need to visualize the corresponding differences between data values
- Avoid the non-monotonic scales such as the rainbow scale
- A sequential scale is also recommended for representing continuous data values
- Helps distinguish higher versus lower values



Median annual income in Texas counties (Source: Figure 4.4 https://clauswilke.com/dataviz/)

Common pitfalls of color use:
lessons

1. Avoid encoding too much or irrelevant information
2. Avoid coloring for the sake of coloring
3. Using monotonic color scales to encode data values
4. **Designing for color-vision deficiency**

Common pitfalls of color use

- Keep in mind that a good proportion of our readers may have some form of color-vision deficiency (e.g., colorblind)

Common pitfalls of color use

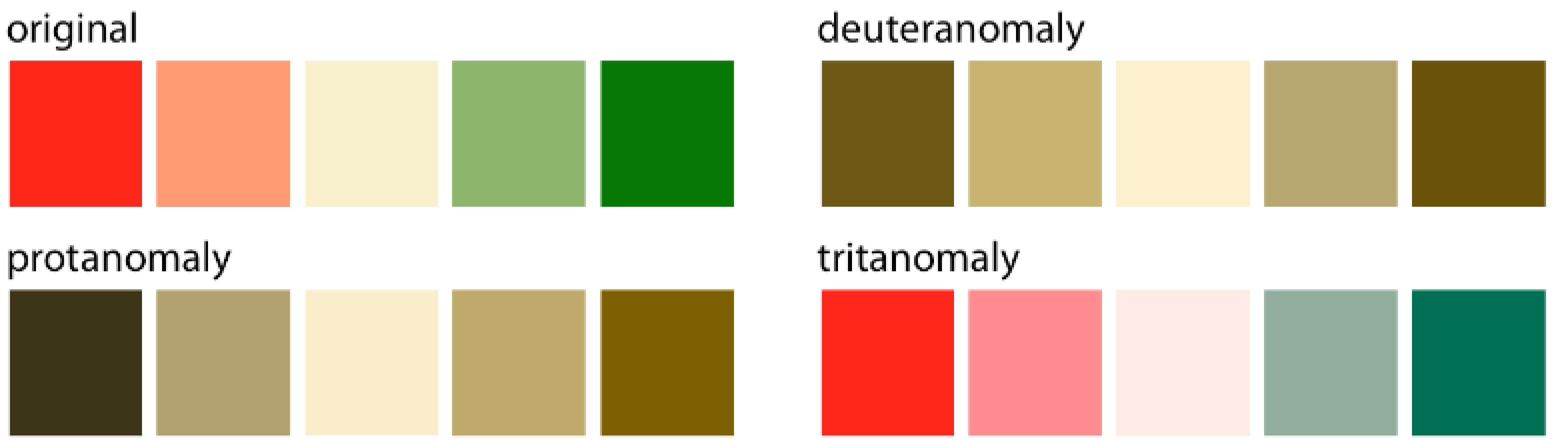
- Keep in mind that a good proportion of our readers may have some form of color-vision deficiency (e.g., colorblind)
- May be difficulty to distinguish certain types of colors, for example red and green or blue and green

Common pitfalls of color use

- Keep in mind that a good proportion of our readers may have some form of color-vision deficiency (e.g., colorblind)
- May be difficulty to distinguish certain types of colors, for example **red** and **green** or **blue** and **green**
- Approximately 8% of males and 0.5% of females suffer from some sort of color-vision deficiency

Common pitfalls of color use

- Red-green deficiency
- deuteranomaly/deuteranopia and protanomaly/protanopia

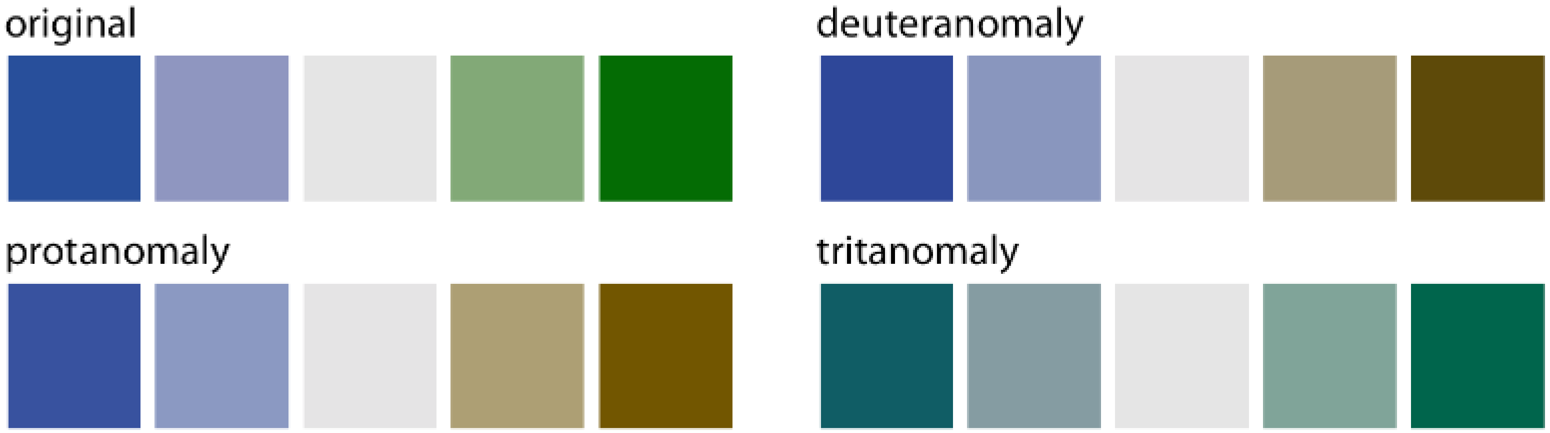


A red–green contrast becomes indistinguishable under red–green cvd (deuteranomaly or protanomaly)

(Source: Figure 19.7 https://clauswilke.com/dataviz/)

Common pitfalls of color use

- Blue-Green deficiency
- tritanomaly/tritanopia

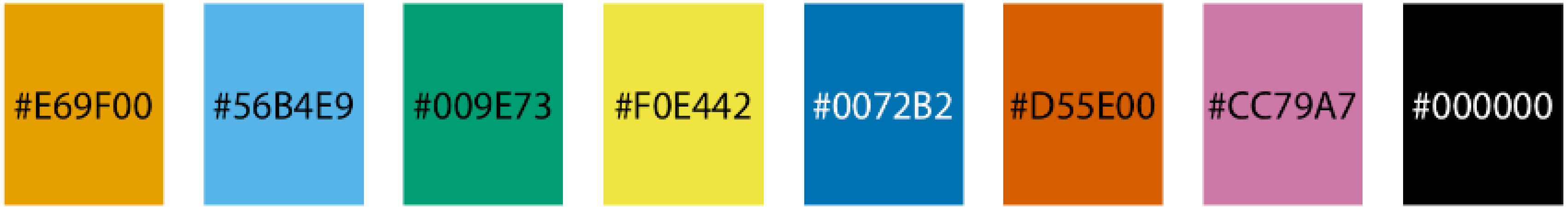


A blue–green contrast becomes indistinguishable under blue–yellow cvd (tritanomaly)

(Source: Figure 19.8 https://clauswilke.com/dataviz/)

Common pitfalls of color use

- Try using qualitative color palette for all color-vision deficiencies



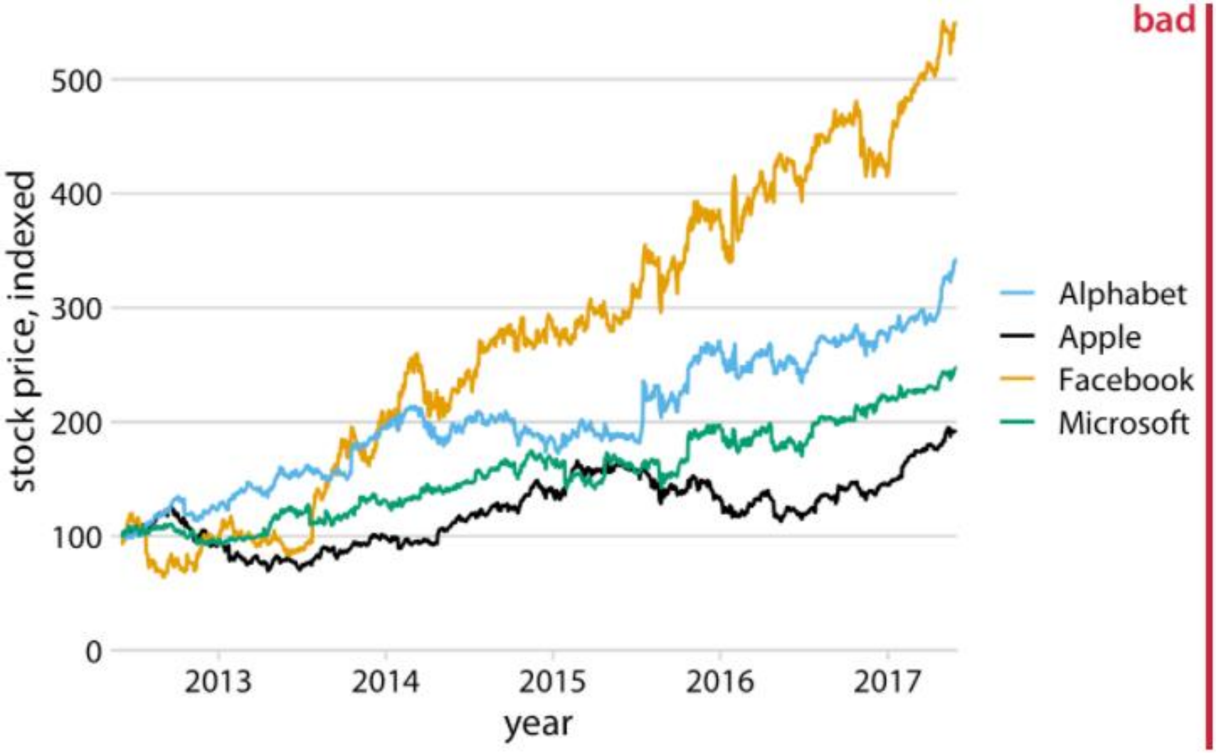
Qualitative color palette for all color-vision deficiencies (Okabe and Ito 2008)

(Source: Figure 19.10 https://clauswilke.com/dataviz/)

Labeling with care

Label

- This figure is labeled as "bad" because it takes considerable mental energy to match the company names in the legend to the data curves.
- Bad for color blinds

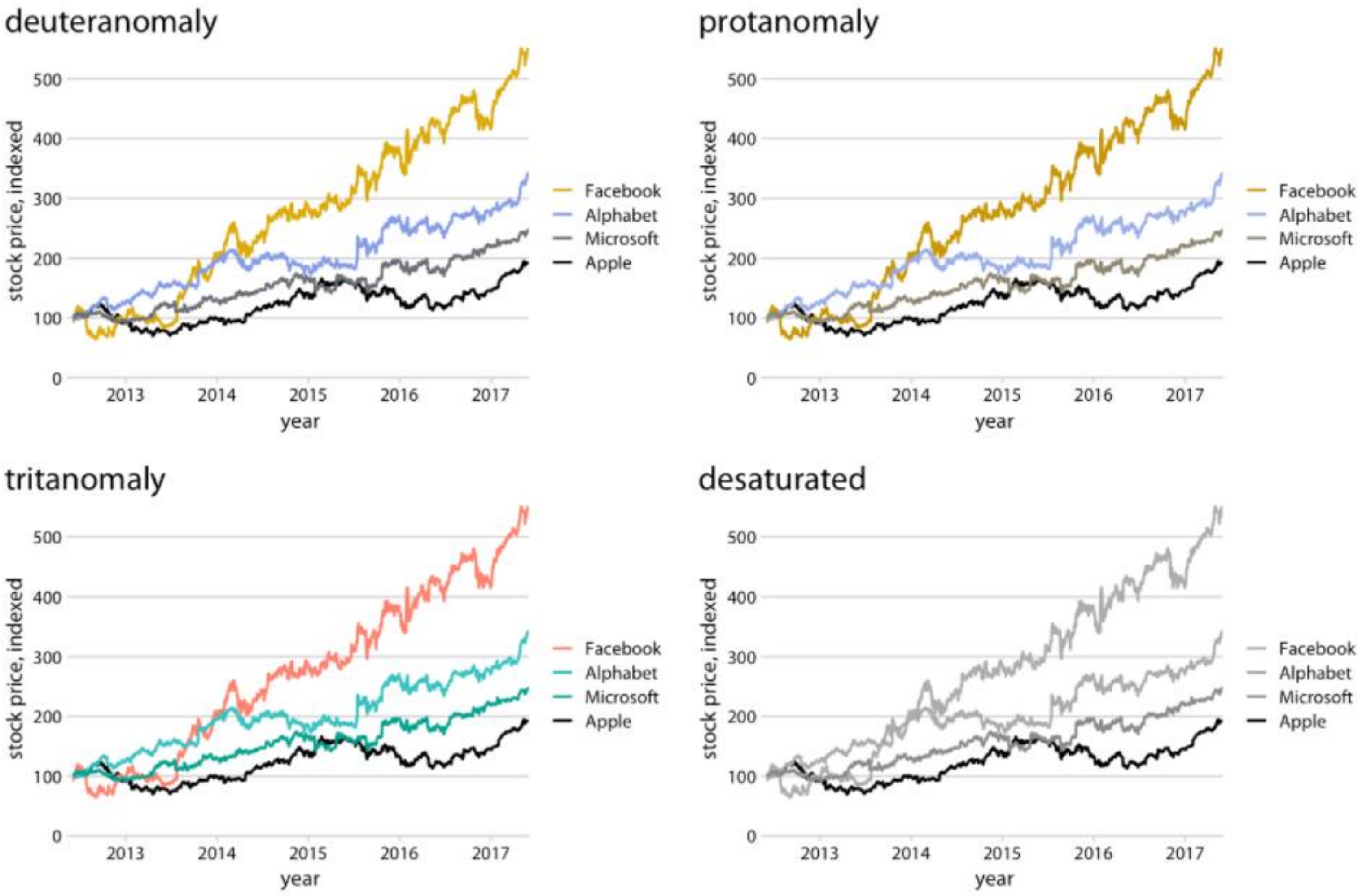


Indexed stock price over time for four major tech companies

(Source: Figure 20.5 https://clauswilke.com/dataviz/)

Label

- Better!
- If there is a clear visual ordering in your data, make sure to match it in the legend.



Color-vision-deficiency simulation

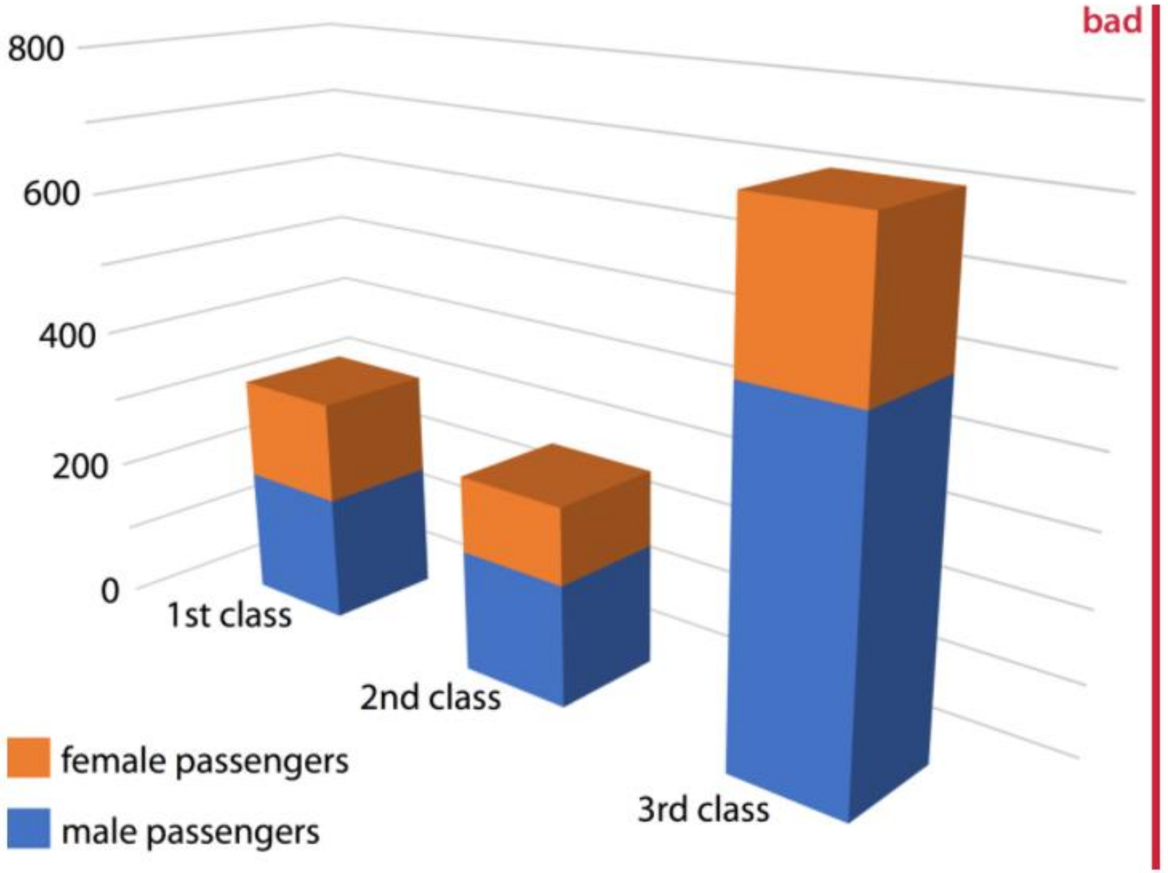
(Source: Figure 20.7 https://clauswilke.com/dataviz/)

Other pitfalls

Don’t go 3D

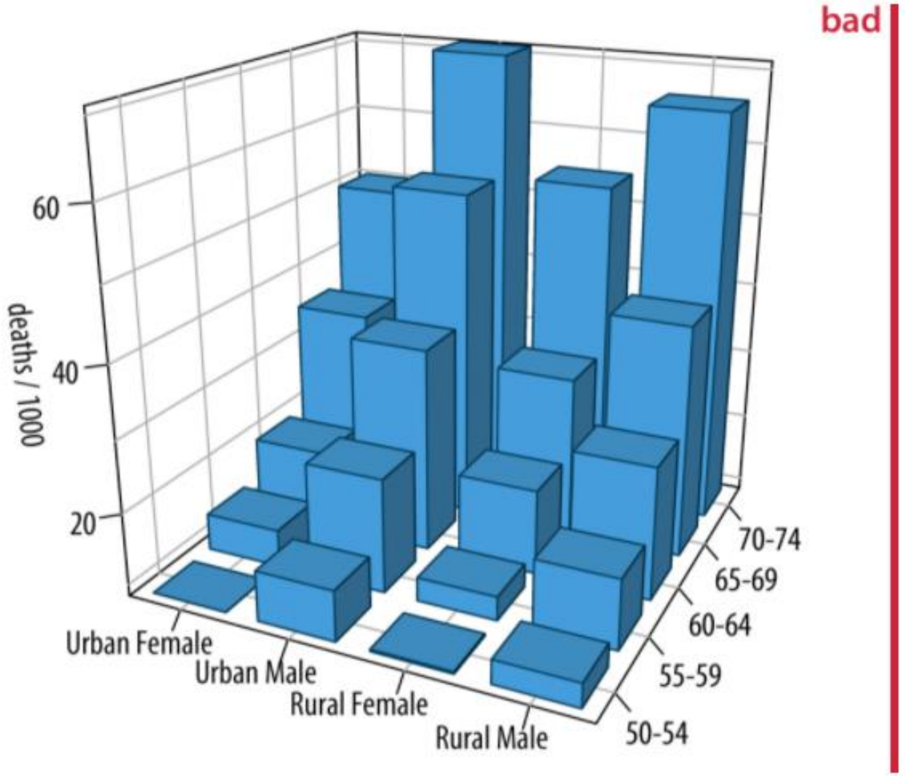
- The problem with gratuitous 3D is that the projection of 3D objects into two dimensions for printing or display on a monitor distorts the data.

- 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.



Don’t go 3D

- The problem with gratuitous 3D 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.

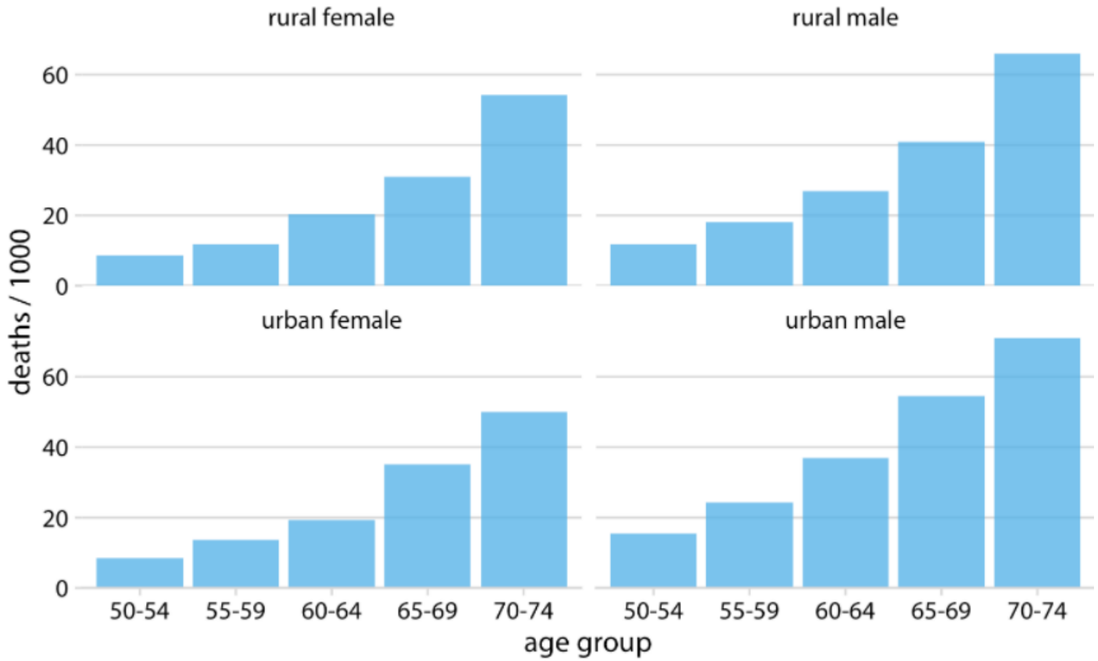


Don’t go 3D

- The problem with gratuitous 3D 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.

- Better!



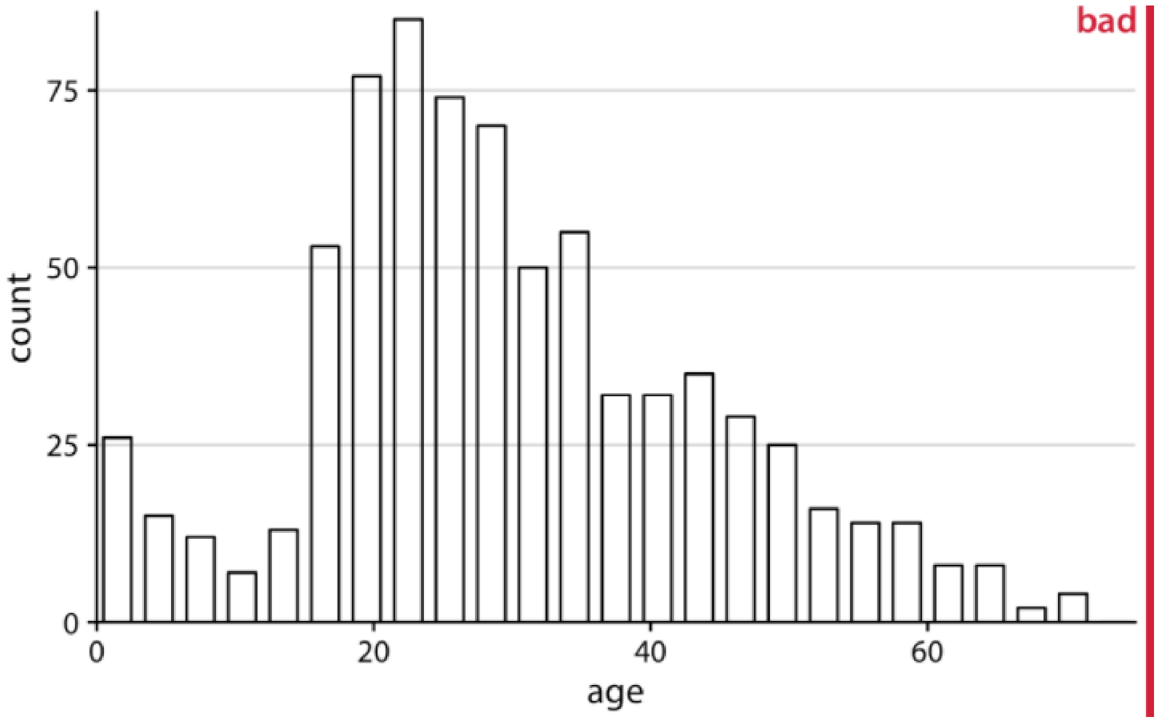
Avoid line drawings

- 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**

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**

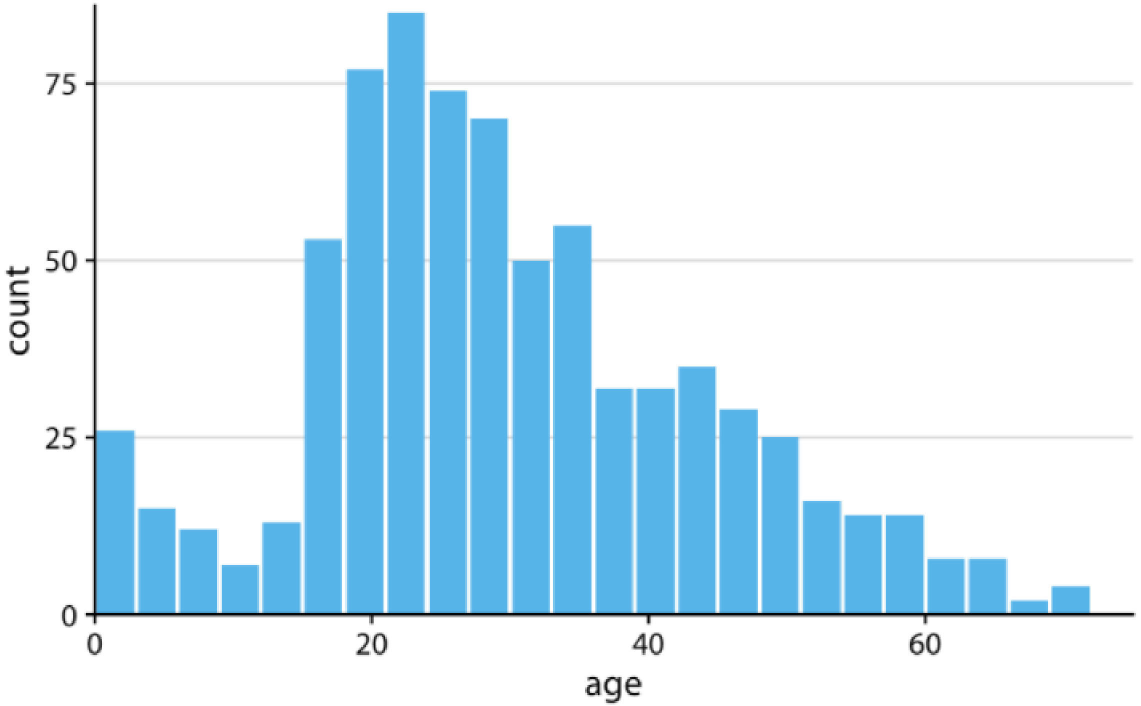


Histogram of the ages of Titanic passengers, drawn with empty bars

(Source: Figure 25.1 https://clauswilke.com/dataviz/)

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
- Better!

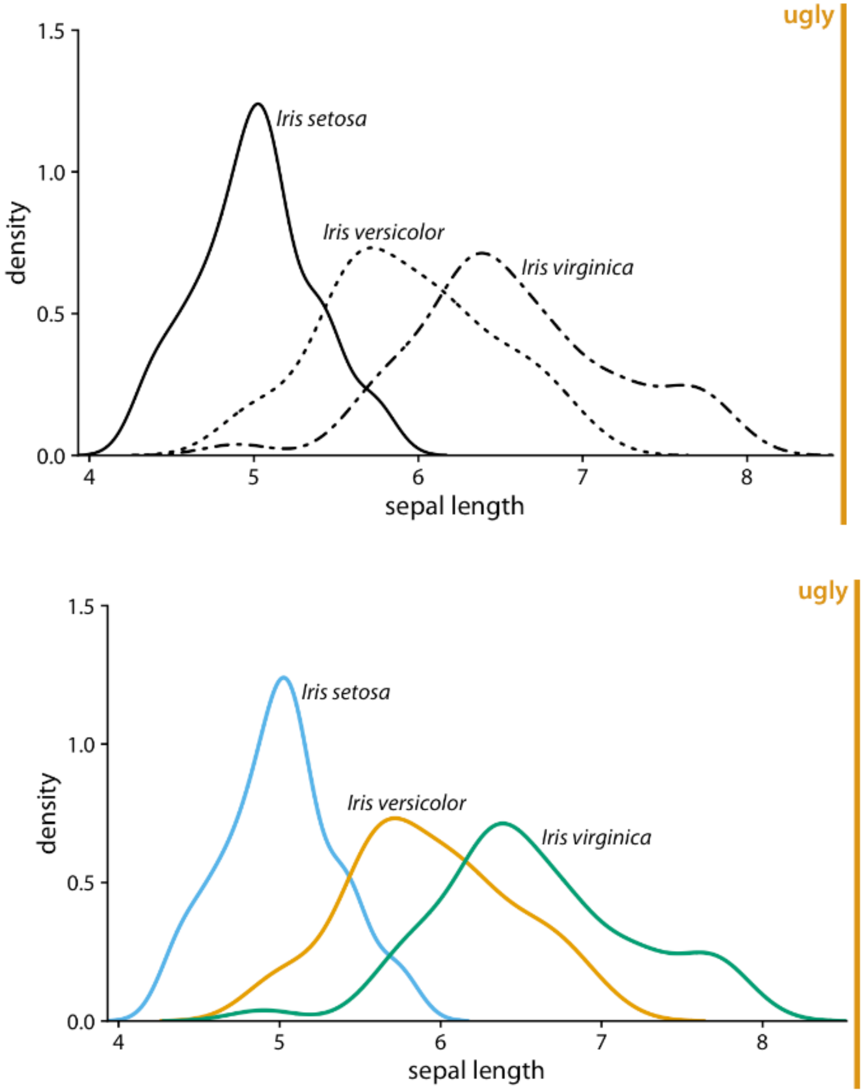


Histogram of the ages of Titanic passengers, drawn with empty bars

(Source: Figure 25.2 https://clauswilke.com/dataviz/)

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



Density estimates of the sepal lengths of three different iris species

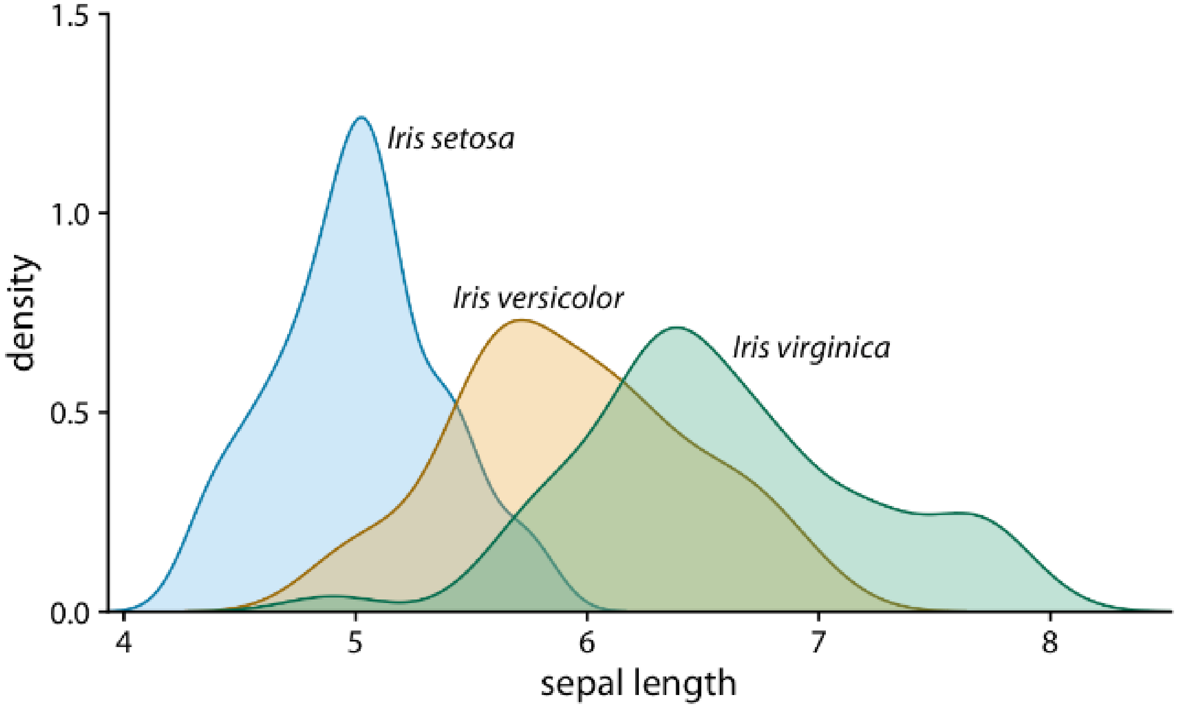
(Source: Figure 25.3-4 https://clauswilke.com/dataviz/)

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

- Better!



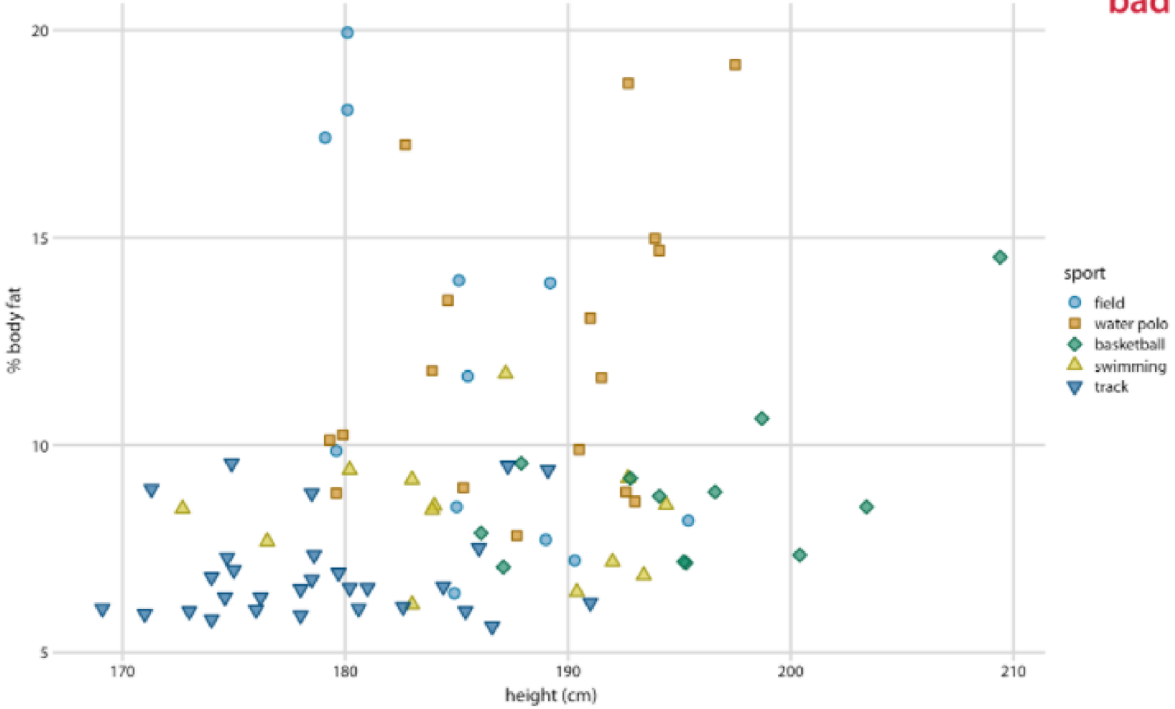
Density estimates of the sepal lengths of three different iris species

(Source: Figure 25.5 https://clauswilke.com/dataviz/)

Use larger axis labels

- Avoid too small to read labels

bad

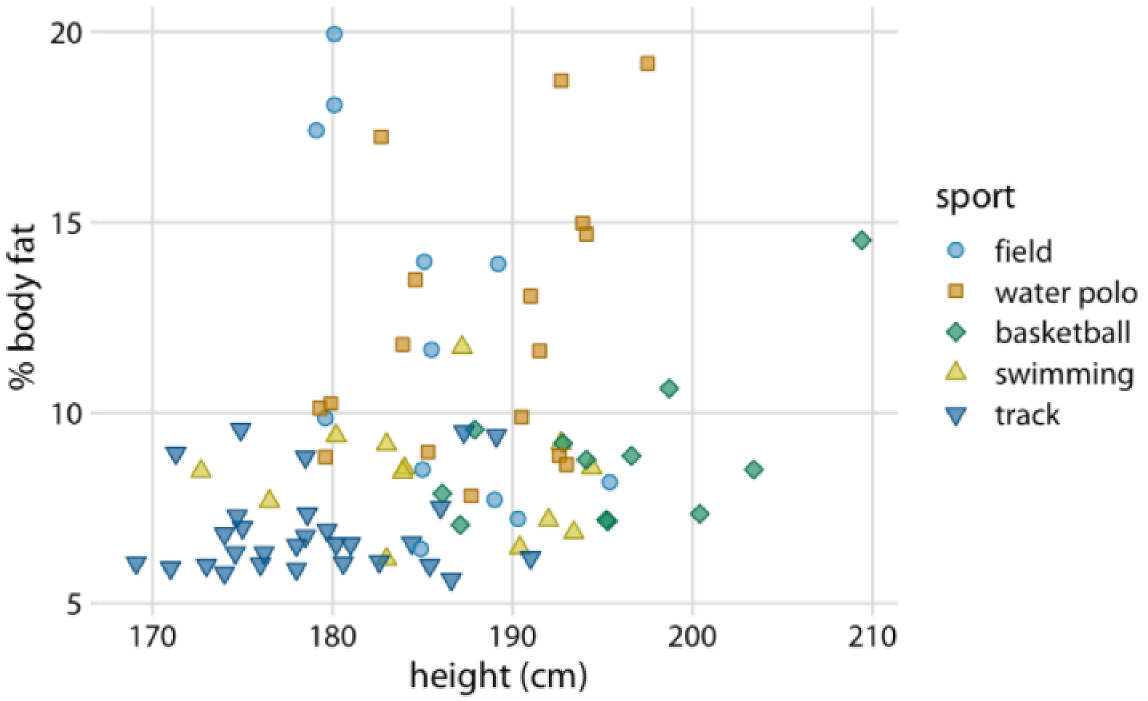


Percent body fat versus height in professional male Australian athletes

(Source: Figure 23.2 https://clauswilke.com/dataviz/)

Use larger axis labels

- Avoid too small to read labels
- Always look at scaled-down versions of your figures to make sure the axis labels are appropriately sized.



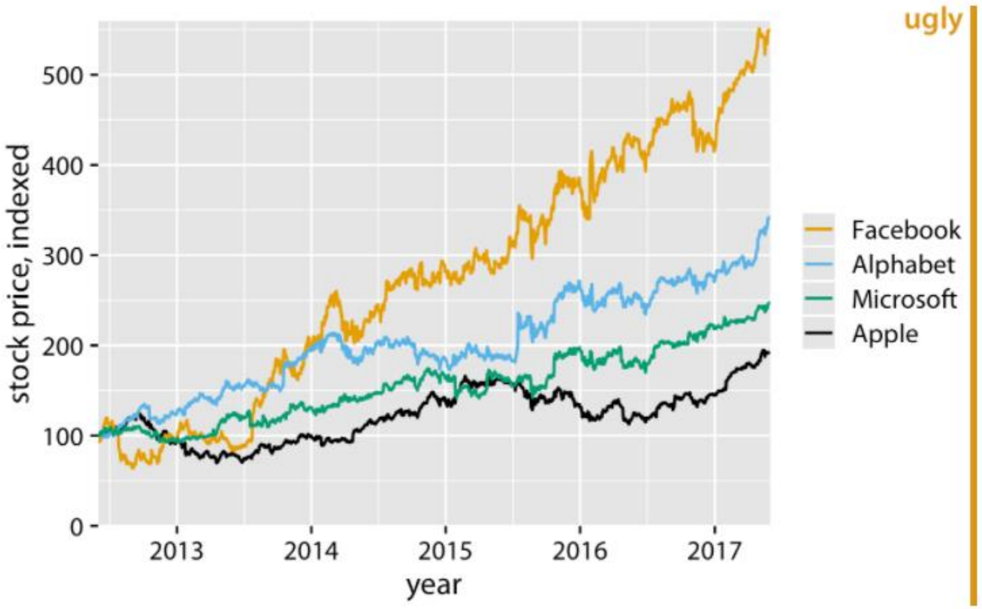
Percent body fat versus height in professional male Australian athletes

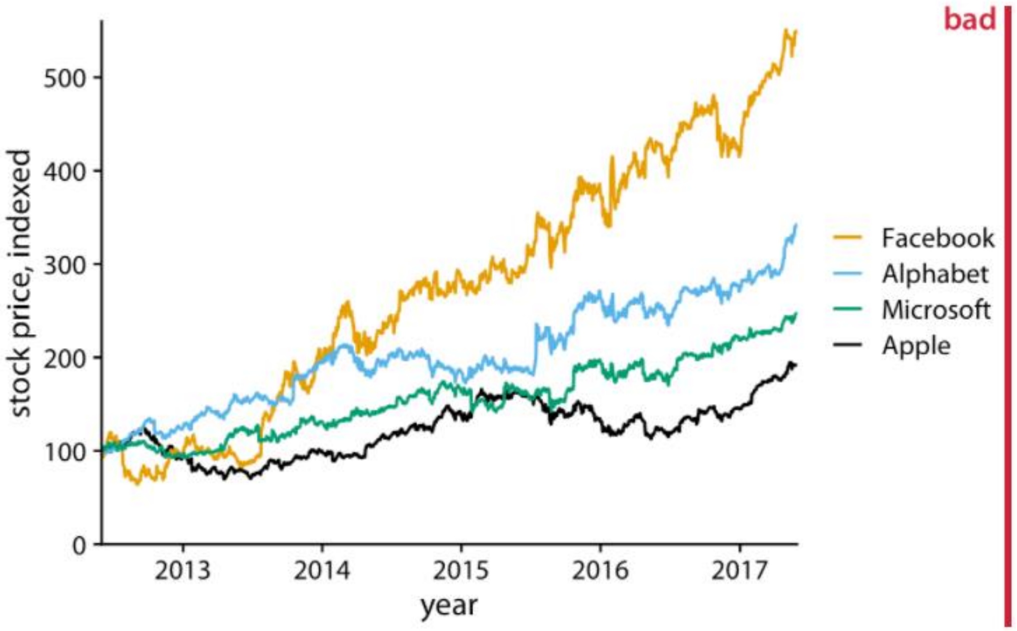
(Source: Figure 23.2 https://clauswilke.com/dataviz/)

Background grids

- **Gridlines** 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, gridlines can add **visual noise**





Stock price over time for four major tech companies

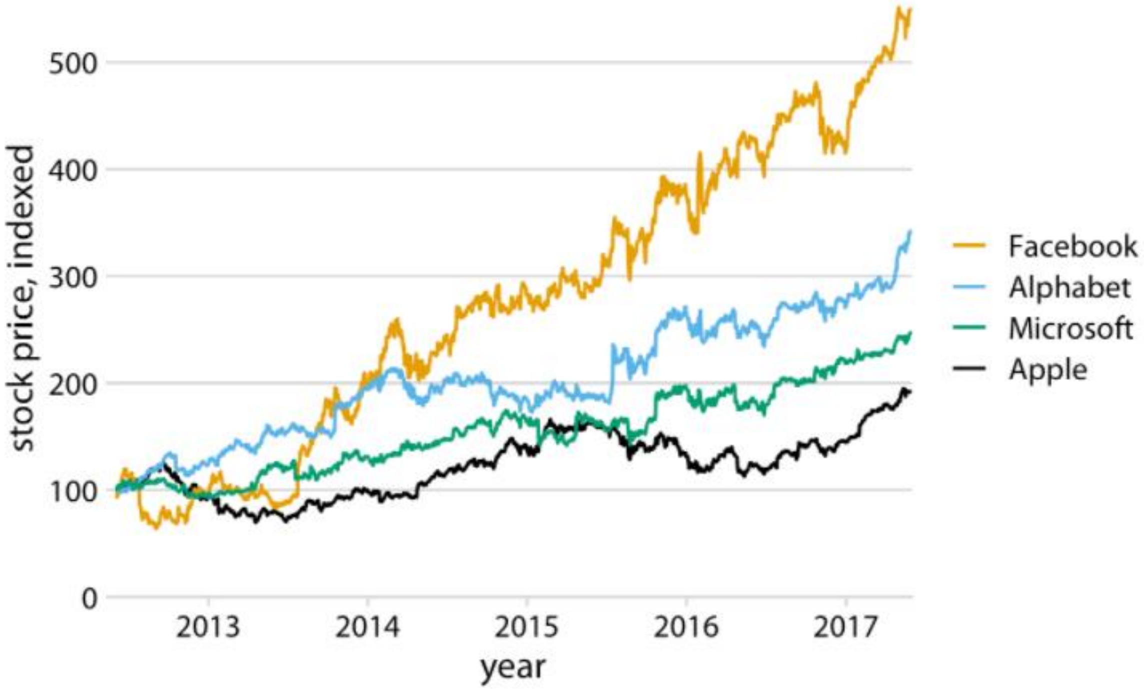
(Source: Figure 23.7 https://clauswilke.com/dataviz/)

Background grids

- **Gridlines** 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, gridlines can add **visual noise**

- Better



Indexed stock price over time for four major tech companies

(Source: Figure 23.9 https://clauswilke.com/dataviz/)

References

- Stone, M., D. Albers Szafir, and V. Setlur. 2014. "An Engineering Model for Color Difference as a Function of Size." In 22nd Color and Imaging Conference. Society for Imaging Science and Technology
- https://www.youtube.com/watch?v=G0BwePFyFIs
- https://clauswilke.com/dataviz/aesthetic-mapping.html
- https://socviz.co/lookatdata.html#lookatdata
- https://www.python-graph-gallery.com/