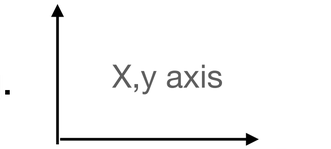
Data Visualization

Procheta Sen

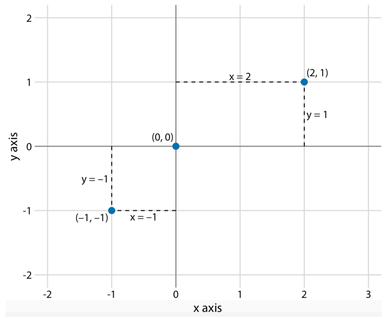
Basic Components of Visualisation

- **Aesthetics**
  - *Position:* Describes where the element is located.
    - E.g. 

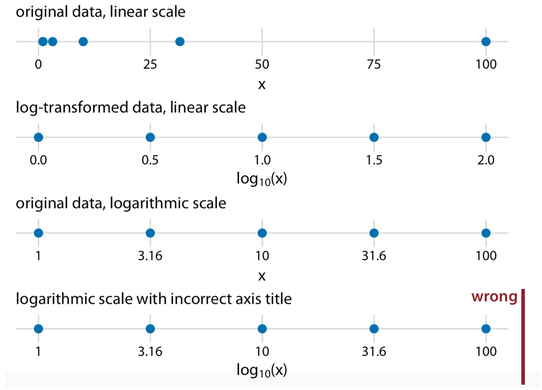
  - *Shape:* 
  - *Size:* 
  - *Color:* 
- **Types of Data**
  - quantitative/numerical continuous, [1,2,3], [1.5,2.6]
  - numerical discrete, categorical (ordered/unordered), [dog, cat, fish], [good, fair, poor]
  - date/time: Jan. 5 2018, 8:03am
  - Text: The quick brown fox jumps over the lazy dog.

Types of Positions

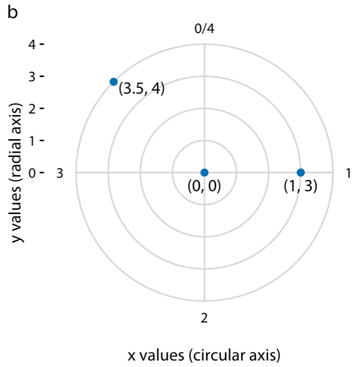
- Cartesian coordinates



- Nonlinear axes

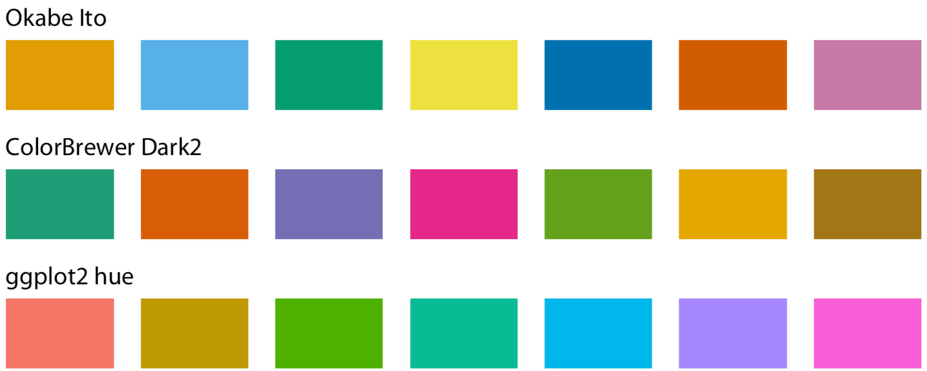


- Coordinate systems with curved axes



Color Scales

- Many appropriate qualitative color scales are readily available.
- ColorBrewer project provides a nice selection of qualitative color scales, including both fairly light and fairly dark colors (Brewer 2017)

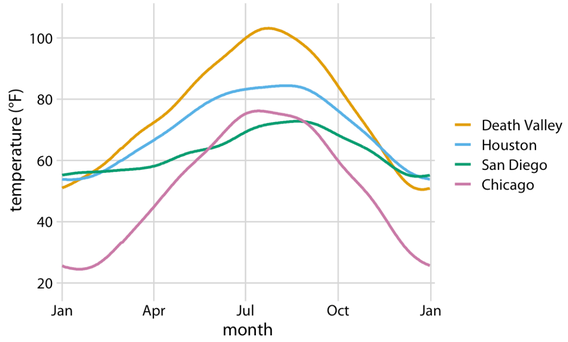


Example qualitative color scales. The Okabe Ito scale is the default scale used throughout this book (Okabe and Ito 2008). The ColorBrewer Dark2 scale is provided by the ColorBrewer project (Brewer 2017). The ggplot2 hue scale is the default qualitative scale in the widely used plotting software ggplot2.

Basic Visualisation Principle

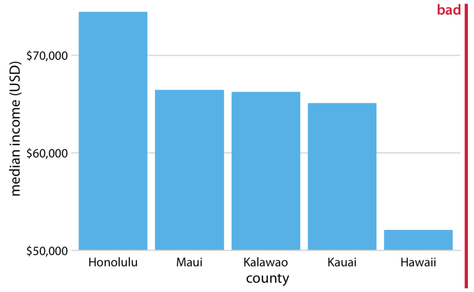
- There is a mapping between data values and aesthetics values. This mapping is done via *scales*.

- A scale must be one-to-one



Daily temperature normals for four selected locations in the U.S. Temperature is mapped to the *y* axis, day of the year to the *x* axis, and location to line color. Data source: NOAA.

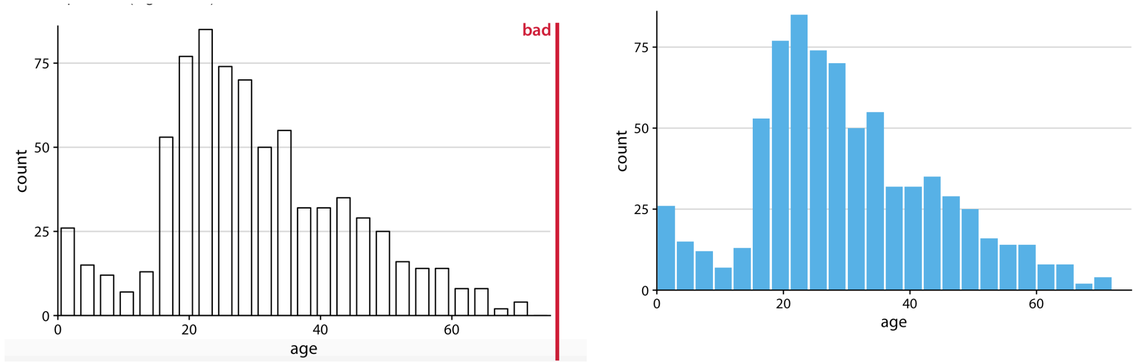
- The sizes of shaded areas in a visualisation need to be proportional to the data values they represent.



Median income in the five counties of the state of Hawaii. This figure is misleading, because the *y* axis scale starts at $50,000 instead of $0. As a result, the bar heights are not proportional to the values shown and the income differential between the county of Hawaii and the other four counties appears much bigger than it actually is. Data source: 2015 Five-Year American Community Survey.

Basic Visualisation Principle (Continued)

- Use larger axis labels.
- Avoid line drawings.

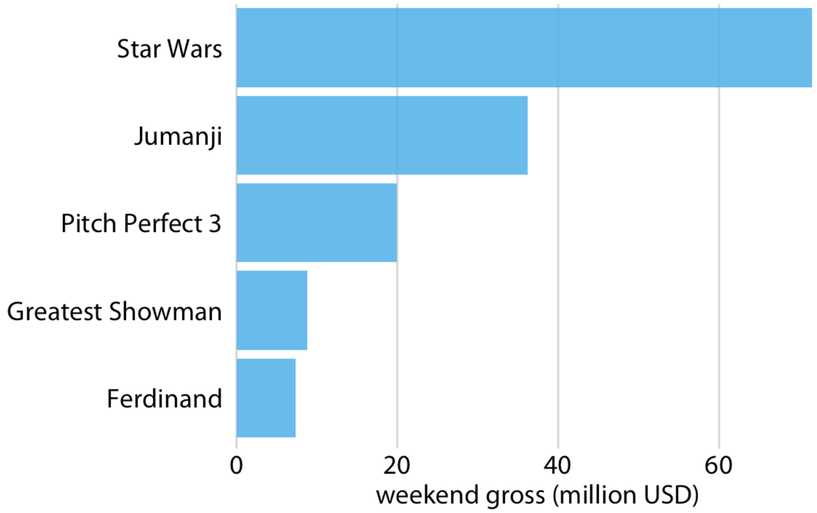


Histogram of the ages of Titanic passengers, drawn with empty bars.

- Choosing the right visualization software.

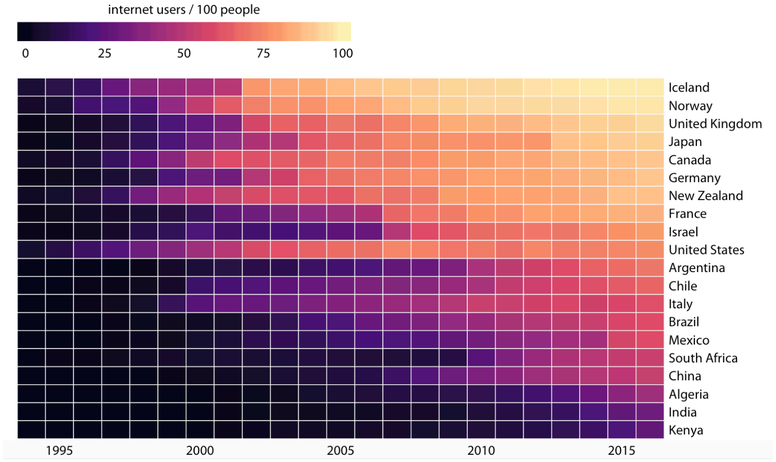
Types of Visualisation: Amount

- Bar Chart



Highest grossing movies for the weekend of December 22-24, 2017, displayed as a horizontal bar plot. Data source: Box Office Mojo (http://www.boxofficemojo.com/).

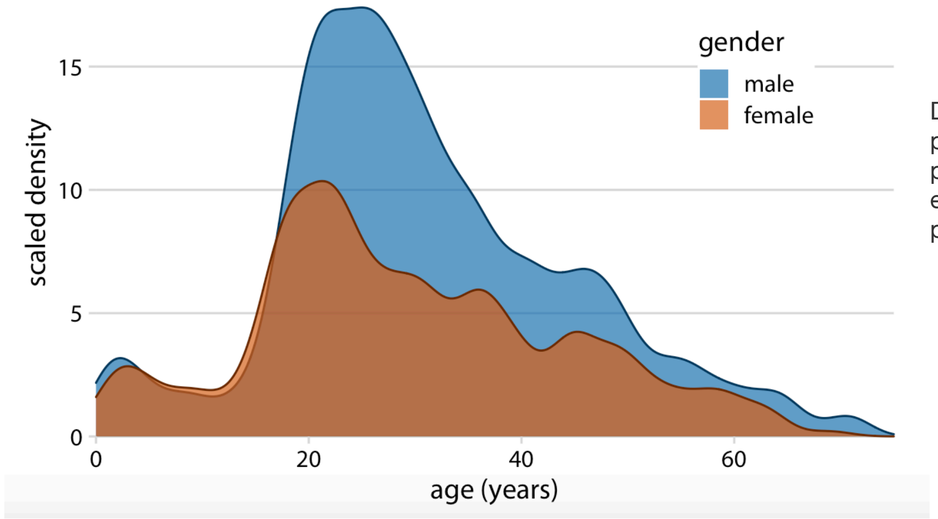
- Heat Map



Internet adoption over time, for select countries. Color represents the percent of internet users for the respective country and year. Countries were ordered by percent internet users in 2016. Data source: World Bank

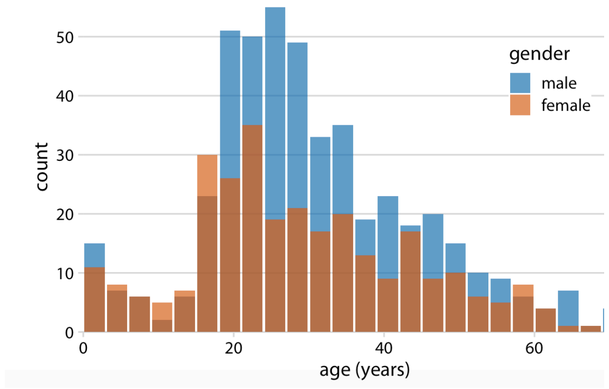
Types of Visualisation: Distribution

- Density Plot



Density estimates of the ages of male and female Titanic passengers. To highlight that there were more male than female passengers, the density curves were scaled such that the area under each curve corresponds to the total number of male and female passengers with known age (468 and 288, respectively).

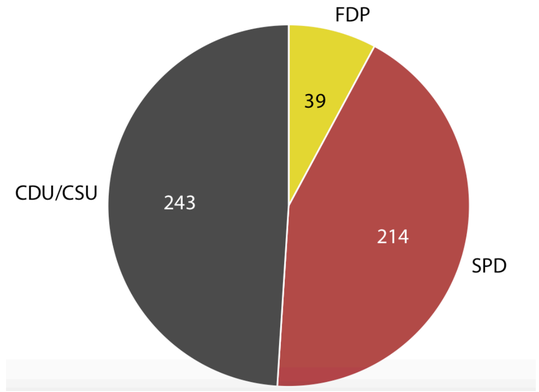
- Histogram



Age distributions of male and female Titanic passengers, shown as two overlapping histograms.

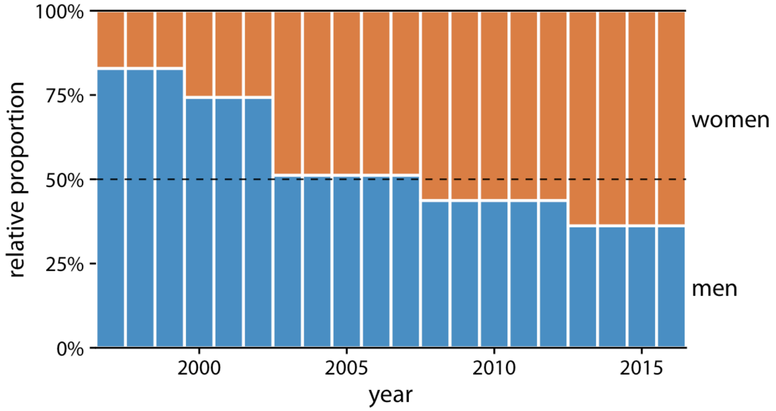
Types of Visualisation: Proportions

- Pie Chart



Party composition of the 8th German Bundestag, 1976–1980, visualized as a pie chart. This visualization shows clearly that the ruling coalition of SPD and FDP had a small majority over the opposition CDU/CSU.

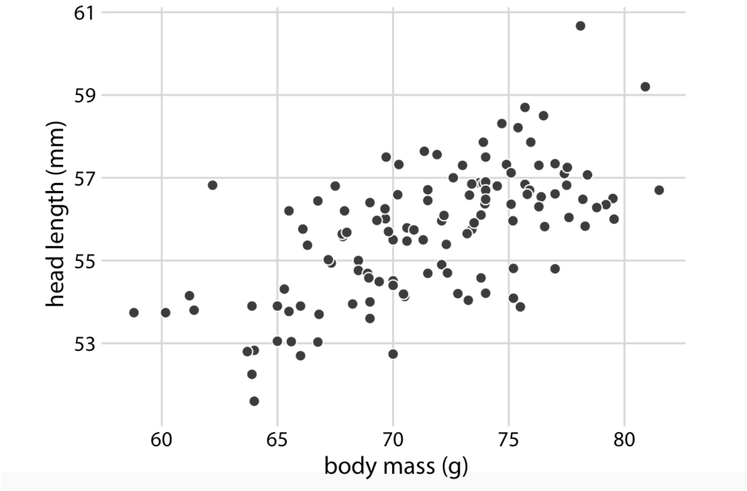
- Side by Side Bars



Change in the gender composition of the Rwandan parliament over time, 1997 to 2016. Data source: Inter-Parliamentary Union (IPU), ipu.org.

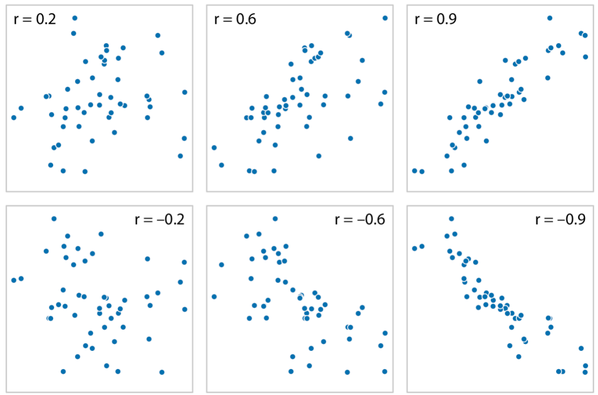
Types of Visualisation: X-Y relationships

- Scatter Plot



Head length (measured from the tip of the bill to the back of the head, in mm) versus body mass (in gram), for 123 blue jays. Each dot corresponds to one bird. There is a moderate tendency for heavier birds to have longer heads. Data source: Keith Tarvin, Oberlin College

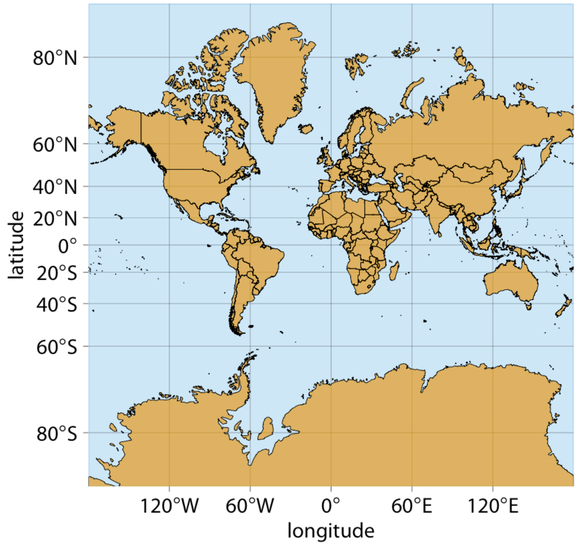
- Correlograms



Examples of correlations of different magnitude and direction, with associated correlation coefficient *r*. In both rows, from left to right correlations go from weak to strong. In the top row the correlations are positive (larger values for one quantity are associated with larger values for the other) and in the bottom row they are negative (larger values for one quantity are associated with smaller values for the other). In all six panels, the sets of *x* and *y* values are identical, but the pairings between individual *x* and *y* values have been reshuffled to generate the specified correlation coefficients.

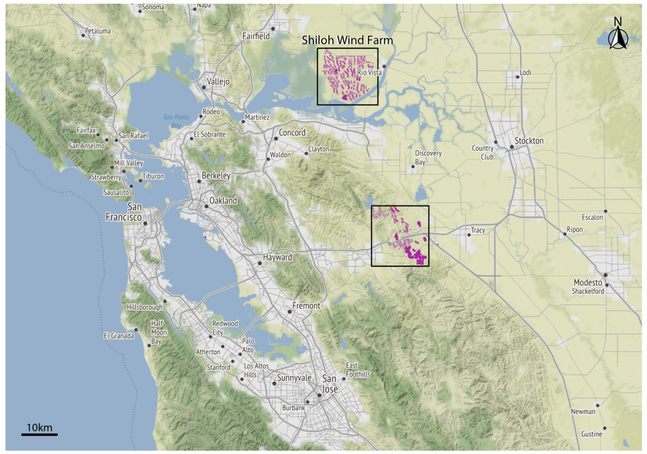
Types of Visualisation: Geospatial Data

- Projections



Mercator projection of the world. In this projection, parallels are straight horizontal lines and meridians are straight vertical lines. It is a conformal projection preserving local angles, but it introduces severe distortions in areas near the poles.

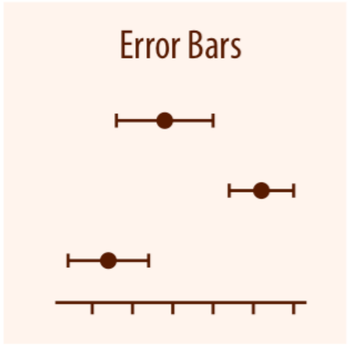
- Layers



Wind turbines in the San Francisco Bay Area. Individual wind turbines are shown as purple-colored dots. Two regions with a high concentration of wind turbines are highlighted with black rectangles. I refer to the wind turbines near Rio Vista collectively as the Shiloh Wind Farm. Map tiles by Stamen Design, under CC BY 3.0. Map data by OpenStreetMap, under ODbL. Wind turbine data: United States Wind Turbine Database

Types of Visualisation: Uncertainty

- Error Bar



- Confidence Strips

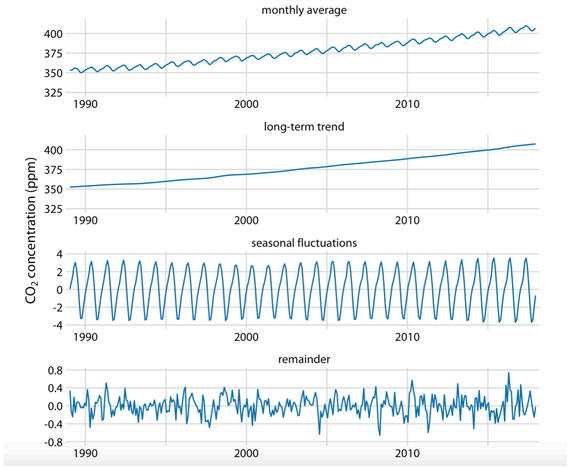
- Confidence Band

Types of Visualisation: Trends

- Smoothing

Daily closing values of the Dow Jones Industrial Average for the year 2009, shown together with their 20-day, 50-day, and 100-day moving averages. (a) The moving averages are plotted at the end of the moving time windows. (b) The moving averages are plotted in the center of the moving time windows. Data source: Yahoo! Finance

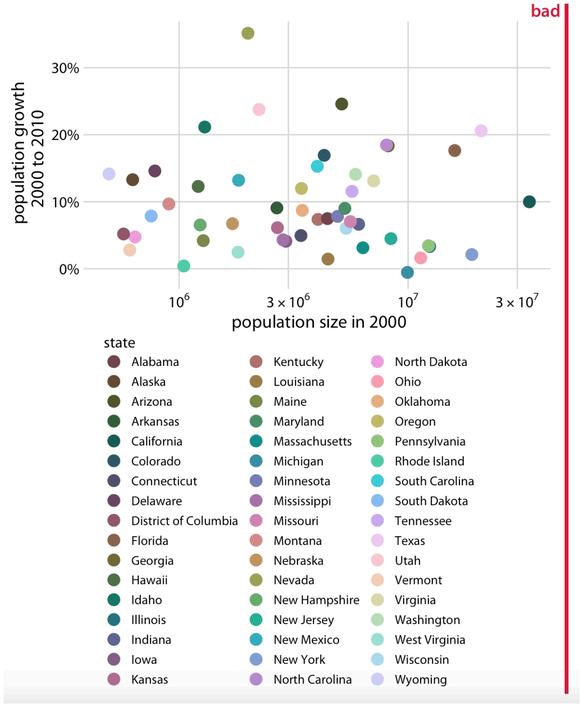
- Time Series



Time-series decomposition of the Keeling curve, showing the monthly average (as in Figure 14.12), the long-term trend, seasonal fluctuations, and the remainder. The remainder is the difference between the actual readings and the sum of the long-term trend and the seasonal fluctuations, and it represents random noise. I have zoomed into the most recent 30 years of data to more clearly show the shape of the annual fluctuations. Data source: Dr. Pieter Tans, NOAA/ESRL, and Dr. Ralph Keeling, Scripps Institution of Oceanography

Common Pitfalls for Color Coding: Part 1

- Encoding too much or irrelevant information
  - As a rule of thumb, qualitative color scales work best when there are three to five different categories that need to be colored. Once we reach eight to ten different categories or more, the task of matching colors to categories becomes too burdensome to be useful, even if the colors remain sufficiently different to be distinguishable in principle.

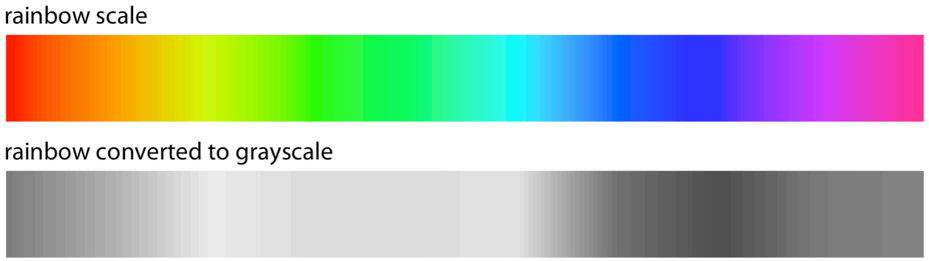


Population growth from 2000 to 2010 versus population size in 2000, for all 50 U.S. states and the Discrict of Columbia. Every state is marked in a different color. Because there are so many states, it is very difficult to match the colors in the legend to the dots in the scatter plot. Data source: U.S. Census Bureau

it is probably best to use color only to indicate the geographic region of each state and to identify individual states by direct labeling, i.e., by placing appropriate text labels adjacent to the data points

Common Pitfalls for Color Coding: Part 2

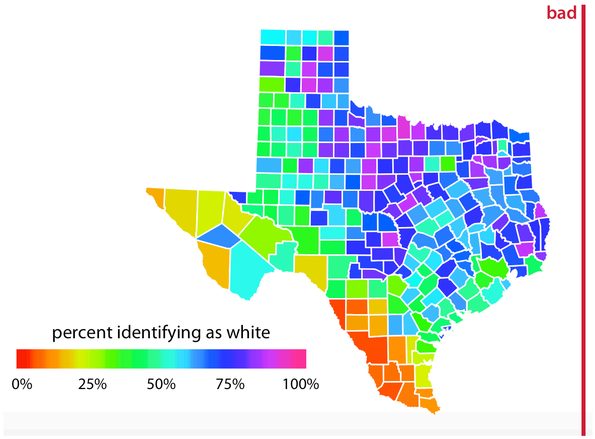
- Using non-monotonic color scales to encode data values



The rainbow colorscale is highly non-monotonic. This becomes clearly visible by converting the colors to gray values. From left to right, the scale goes from moderately dark to light to very dark and back to moderately dark. In addition, the changes in lightness are very non-uniform. The lightest part of the scale (corresponding to the colors yellow, light green, and cyan) takes up almost a third of the entire scale while the darkest part (corresponding to dark blue) is concentrated in a narrow region of the scale.

Common Pitfalls for Color Coding: Part 3

- Not designing for color-vision deficiency



Percentage of people identifying as white in Texas counties. The rainbow color scale is not an appropriate scale to visualize continuous data values, because it tends to place emphasis on arbitrary features of the data. Here, it emphasizes counties in which approximately 75% of the population identify as white. Data source: 2010 Decennial U.S. Census

Python Visualisation Library

- Matplotlib
- Seaborn
- Plotting
- Bokeh
- Pygal
- Geoplotlib