VSB TECHNICAL
UNIVERSITY
OF OSTRAVA

FACULTY OF ELECTRICAL
ENGINEERING AND COMPUTER
SCIENCE

DEPARTMENT
OF COMPUTER
SCIENCE

# Data Visualization

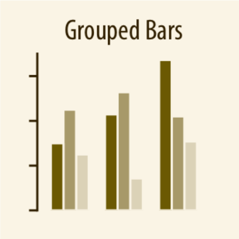
460-4120

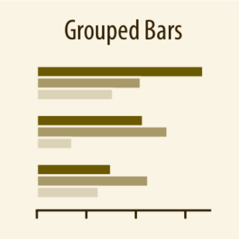
Fall 2024

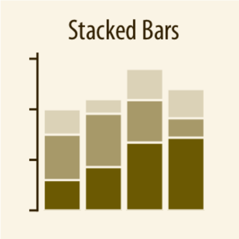
Last update 2. 10. 2024

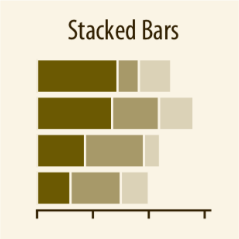
# Charts Overview

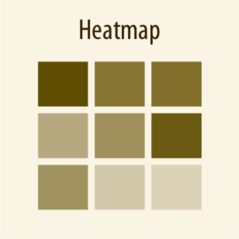
- Amounts (numerical values shown for some set(s) of categories)
  - Bar charts (grouped or stacked when two or more sets of categories), heat maps (amount via color)







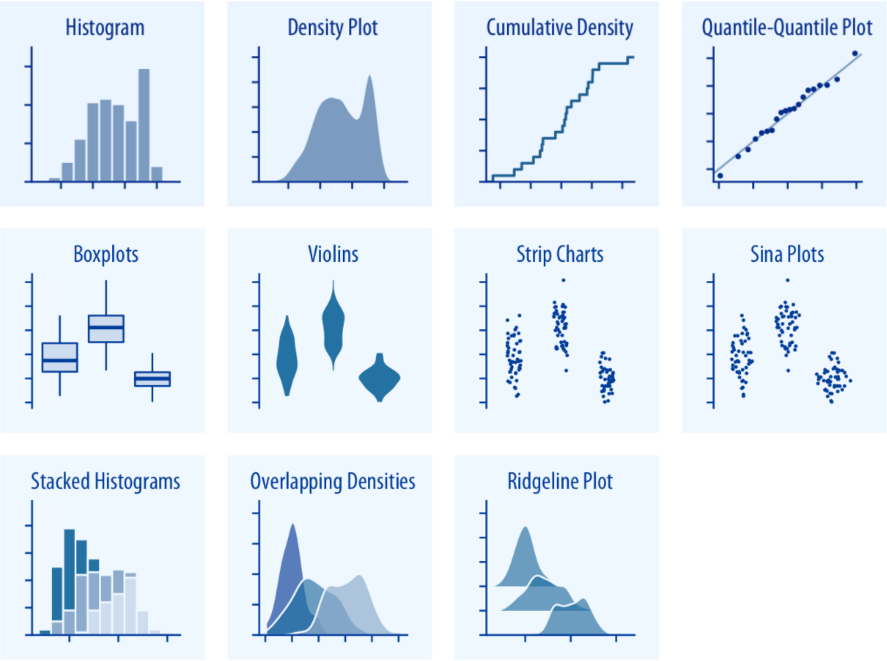




Source: Claus O. Wilke, Fundamentals of Data Visualization

# Charts Overview

- Distributions (one or more distributions or changes in distributions)
  - Histograms, density plots, box plots, violin plots, ridgeline plots

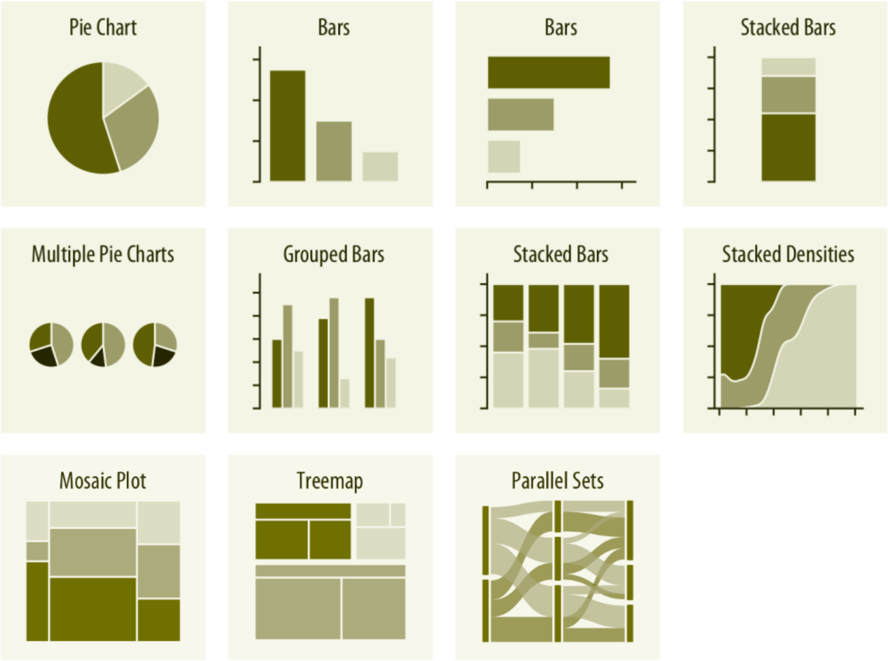


A Q–Q plot is used to compare the shapes of two distributions. If the two distributions are similar, the points in the Q–Q plot will approximately lie on the identity line.

Source: Claus O. Wilke, Fundamentals of Data Visualization

# Charts Overview

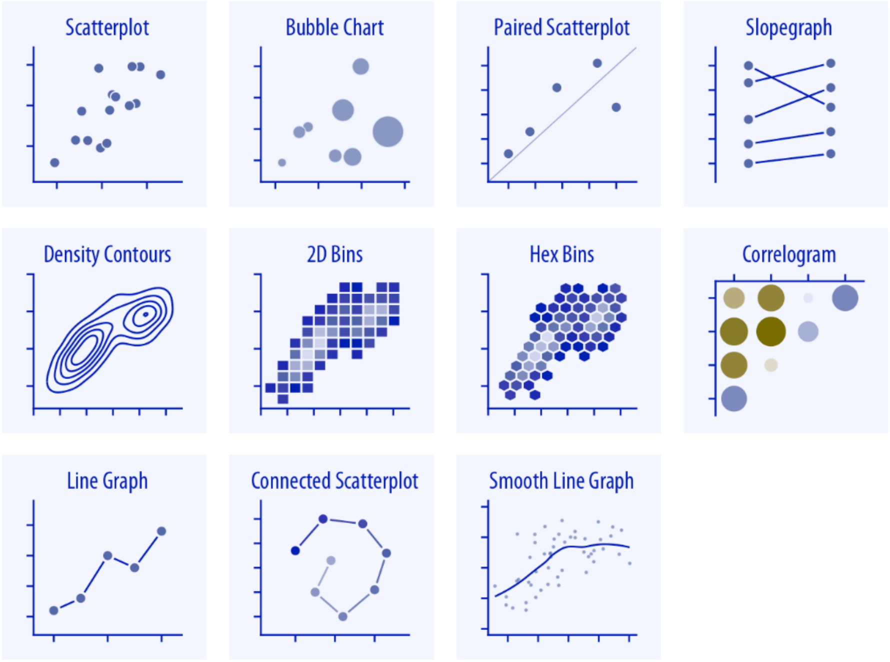
- Proportions (emphasize individual parts of a whole and highlight fractions)
    - Pie chart, bar charts (grouped or stacked), mosaic plots, treemaps, parallel sets



Source: Claus O. Wilke, Fundamentals of Data Visualization

# Charts Overview

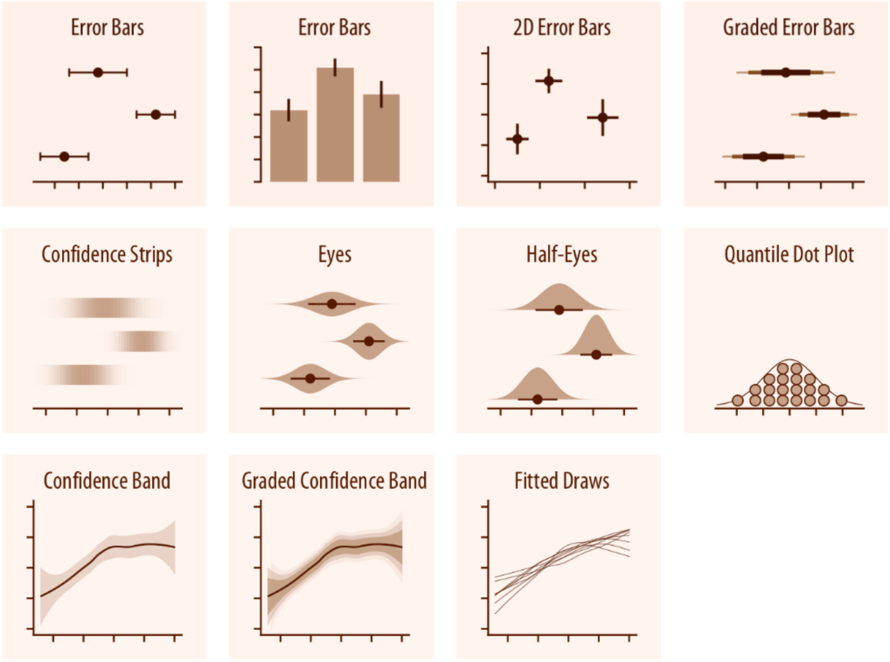
- x-y relationships (show one quantitative variable relative to another)
  - Scatter plot, slopegraph, density contours, 2D bins, correlograms, line graphs



Source: Claus O. Wilke, Fundamentals of Data Visualization

# Charts Overview

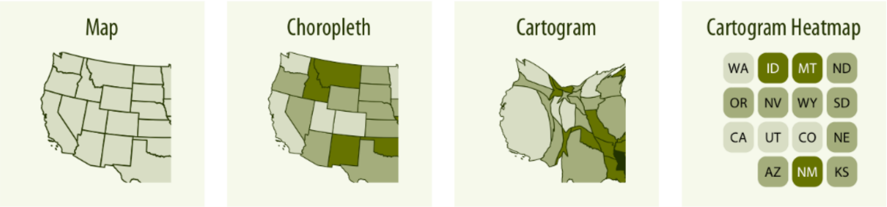
- Uncertainty (indicate the range of likely values for some estimate or measurement, visualize the actual confidence or posterior distributions)
  - Error bars, eyes plots, confidence band



Source: Claus O. Wilke, Fundamentals of Data Visualization

# Charts Overview

- Geospatial data (emphasize individual parts of a whole and highlight fractions)
  - Maps, choropleths, cartograms



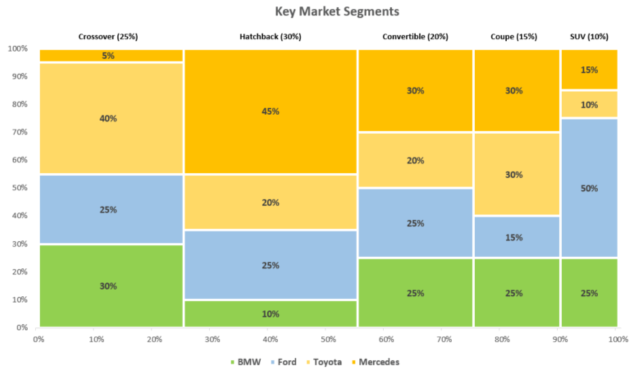
Source: Claus O. Wilke,
Fundamentals of Data
Visualization

# Basic Charts

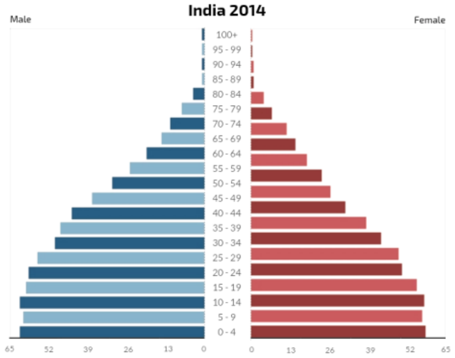
- Point or line charts
  - Illustrate trends in data over a period of time or a particular correlation
  - Each value is plotted on the graph as a vertex of a polyline (or other interpolation curve) over the period of time
- Bar charts
  - Straightforward way to compare various categories
  - One axis features the categories being compared, while the other axis represents the value of each
- Pie charts
  - Efficient visual tool for comparing parts of a whole

# Basic Charts

- A mosaic (Mekko) charts
  - Used to compare multiple variables or multiple categories at the same time

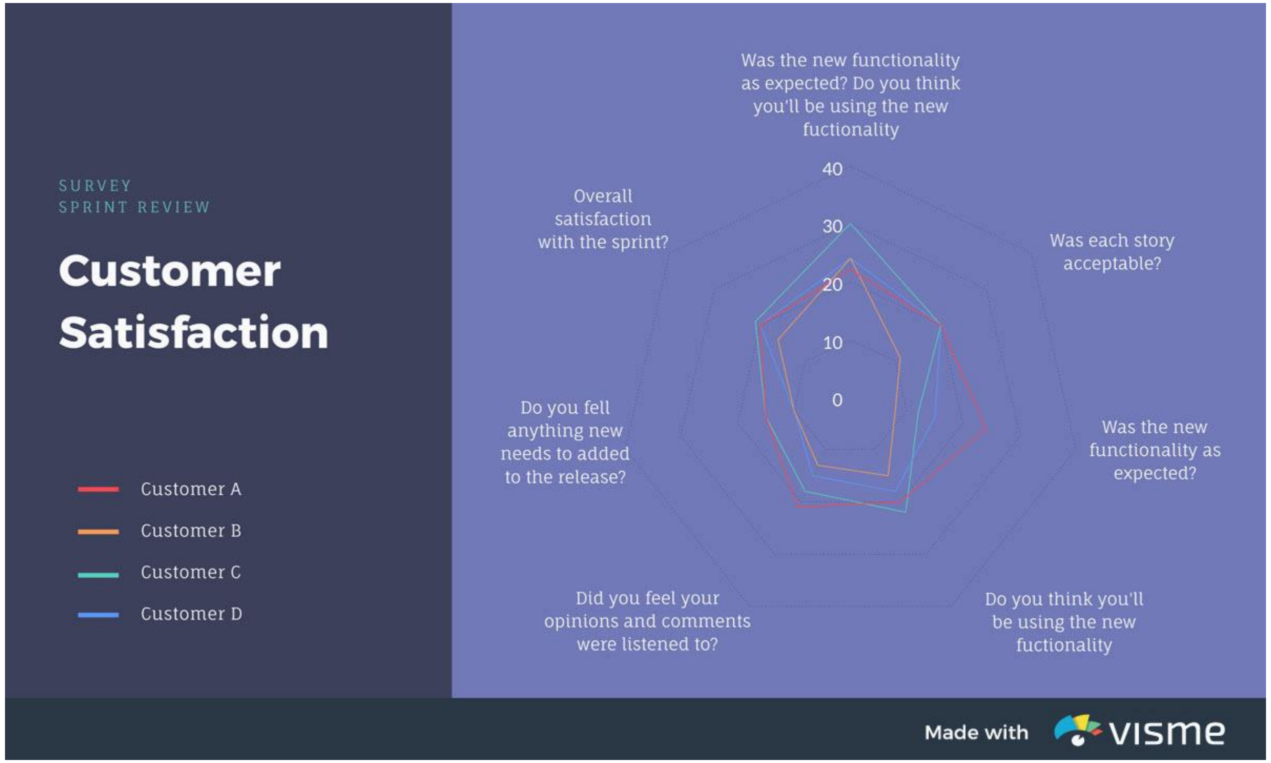


- Population pyramids
  - Used to visually compare two groups



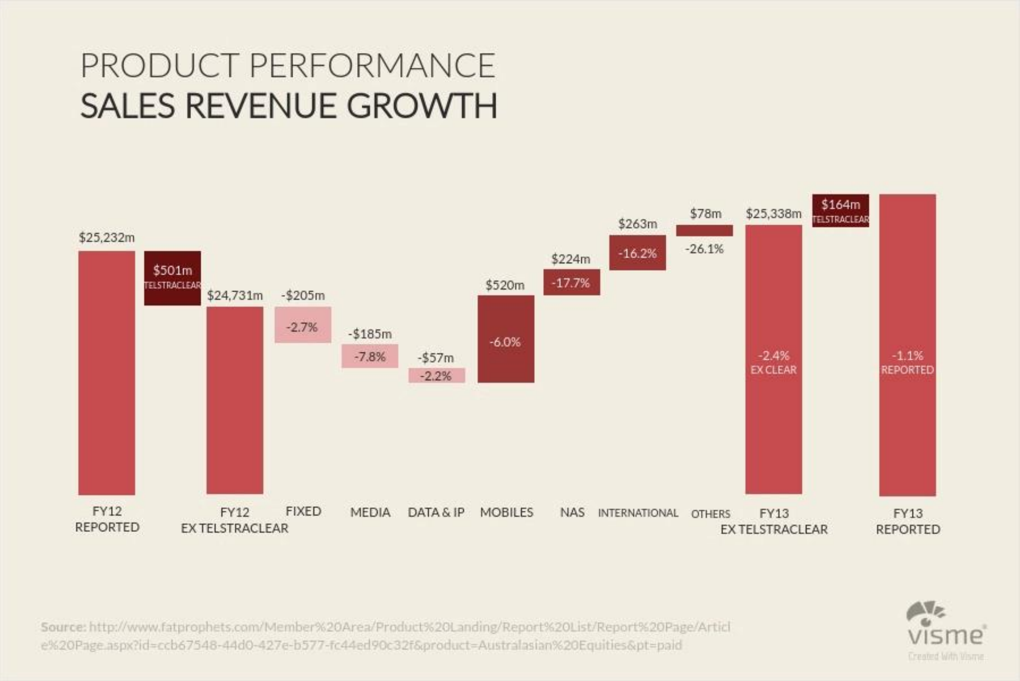
# Basic Charts

- Spider (radar, star) charts
  - Used to visually compare three or more quantitative variables

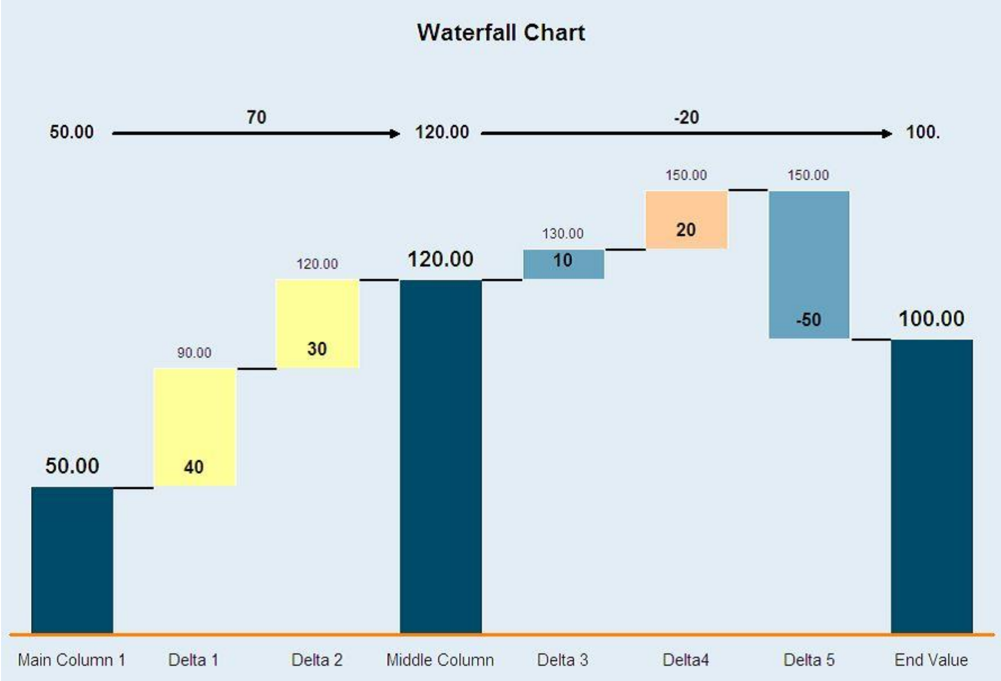


# Basic Charts

- Waterfall (flying brick, bridge, Mario) charts
    - Helps in understanding the cumulative effect of sequentially introduced positive or negative values

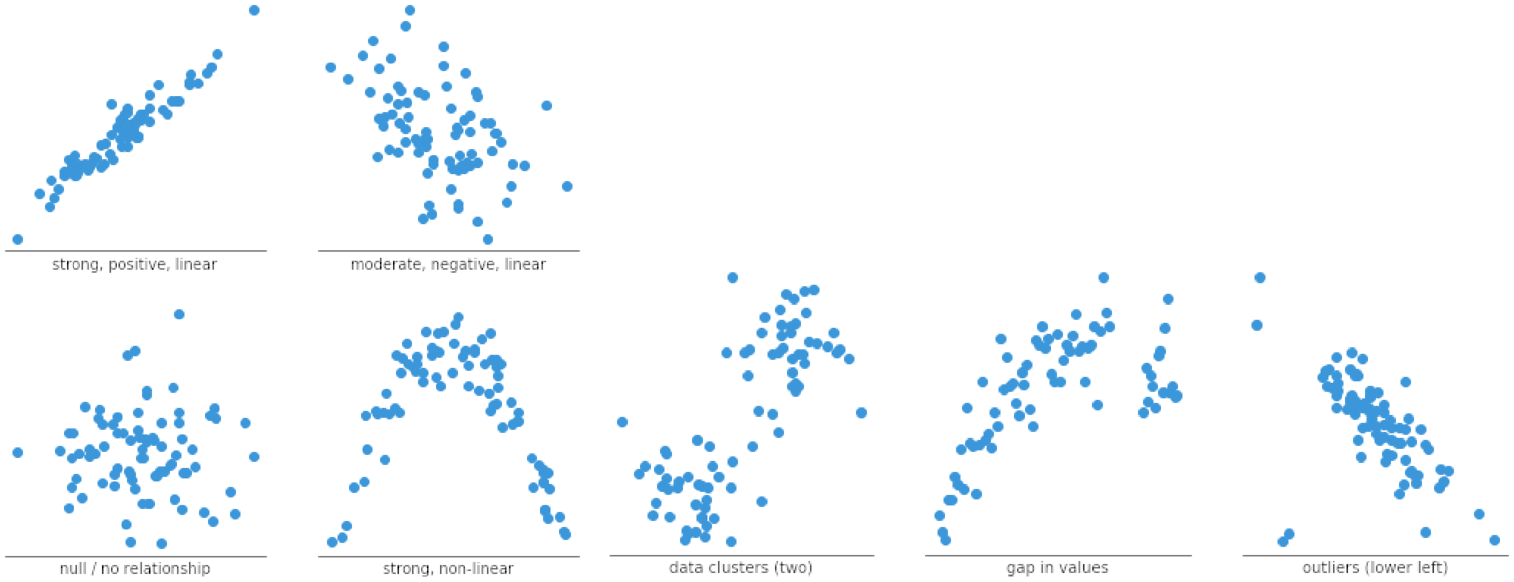


Source: http://www.fatprophets.com/Member%20Area/Product%20Landing/Report%20List/Report%20Page/Article%20Page.aspx?id=ccb67548-44d0-427e-b577-fc44ed90c32f&product=Australasian%20Equities&pt=paid



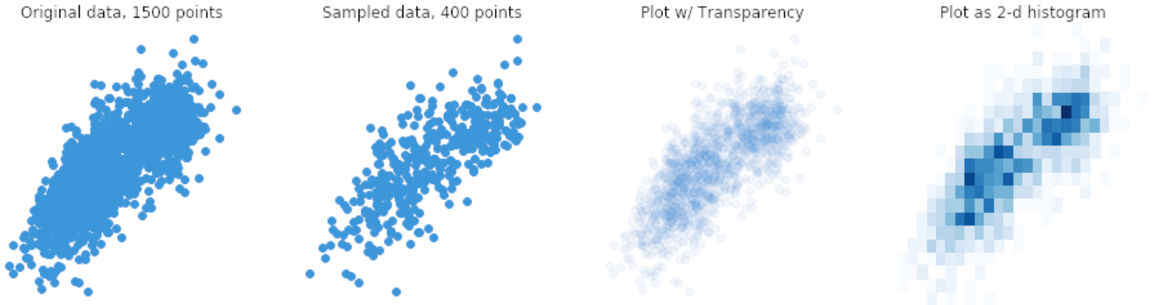
# Basic Charts

- Scatter plots (scattergrams)
  - Uses dots to represent values for two different numeric variables
  - Scatter plots are used to observe relationships between variables



# Basic Charts

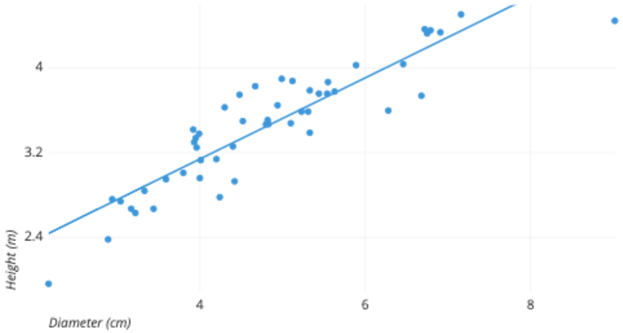
- Scatter plots (scattergrams)
  - Common issue related with scatter plots is overplotting
  - It can be difficult to tell how densely packed data points are



  - Be aware of that "correlation does not imply causation" – the inability to legitimately deduce a cause-and-effect relationship between two events or variables solely on the basis of an observed association or correlation between them

# Basic Charts

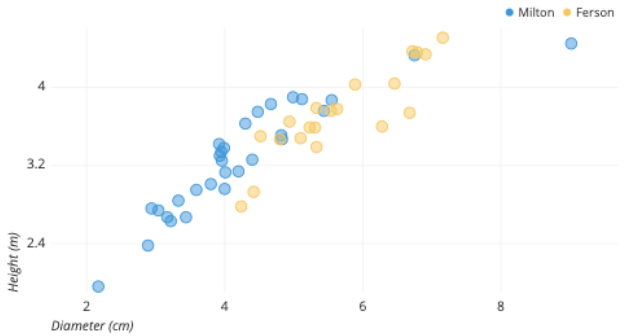
- Scatter plots (scattergrams)
  - Add a trend line (or any other appropriate curve) to show how strong the relationship between the two variables is

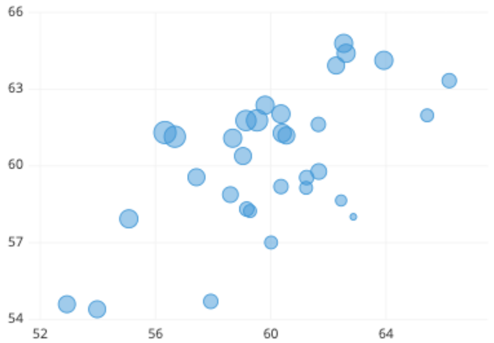


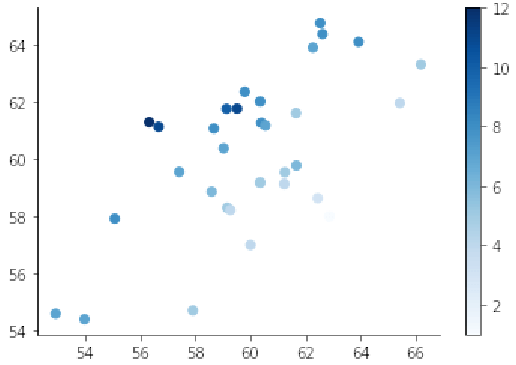
  - It will also reveal any unusual points that are affecting the shape of the trend line

# Basic Charts

- Scatter plots (scattergrams)
  - Add a categorical third variable encoded by modifying how the points are plotted (color, shape) or add numerical variable via hue, size



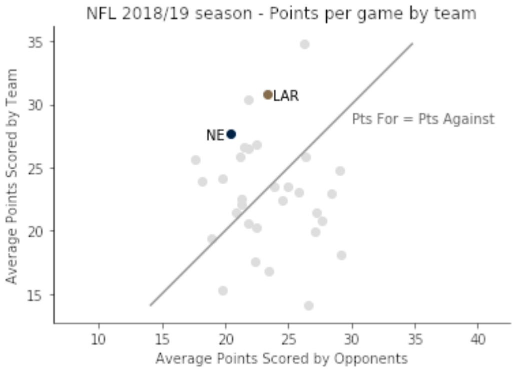


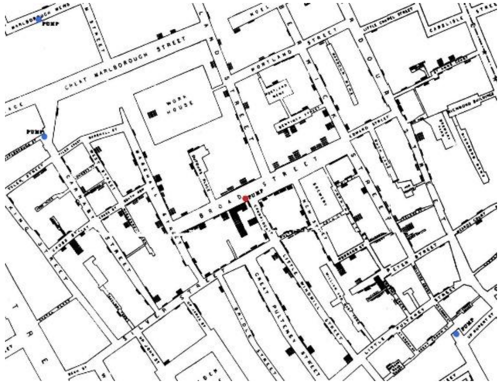


  - It will also reveal membership of each point to a respective/potential group

# Basic Charts

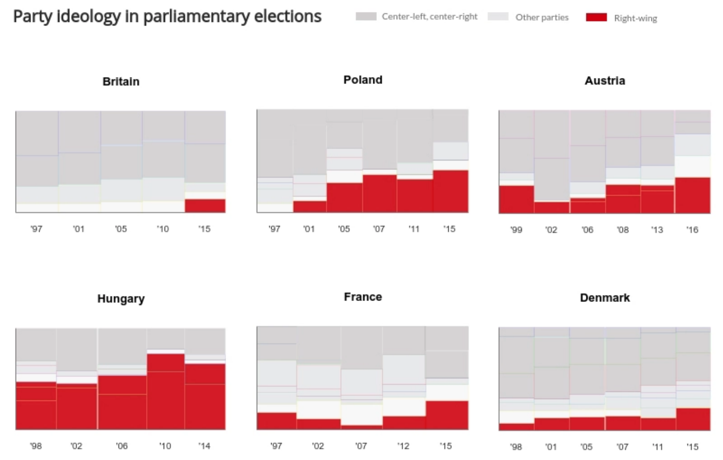
- Scatter plots (scattergrams)
  - To present insights, highlight particular points of interest through the use of annotations and color and desaturate remaining points
  - Additional context can be provided by a background (scatter maps)



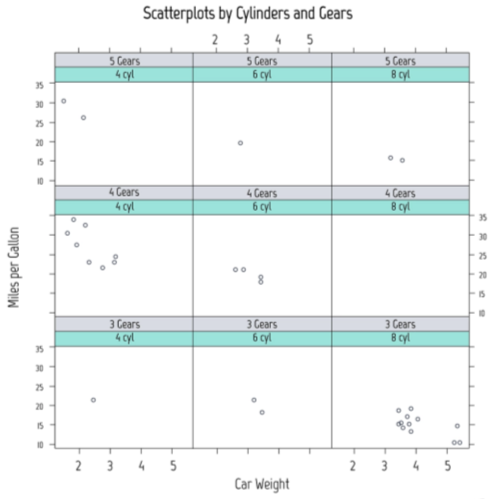


# Basic Charts

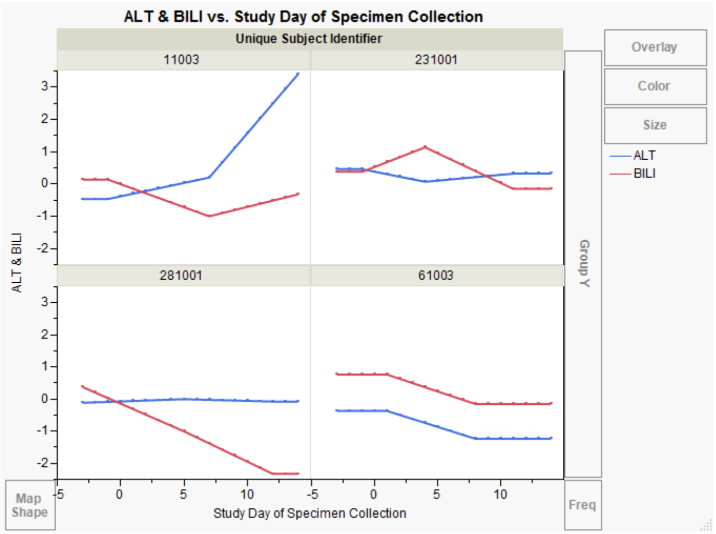
- Trellis plot (lattice or panel plot)
  - A group of small plots (bar charts, scatter plots, line graphs) arranged in a grid
  - Each plot represent a different condition or item but all plots share the same scale
  - Make complex or high-dimensional data easier to interpret



Source: https://www.nytimes.com/interactive/2016/05/22/world/europe/europe-right-wing-austria-hungary.html?_r=1

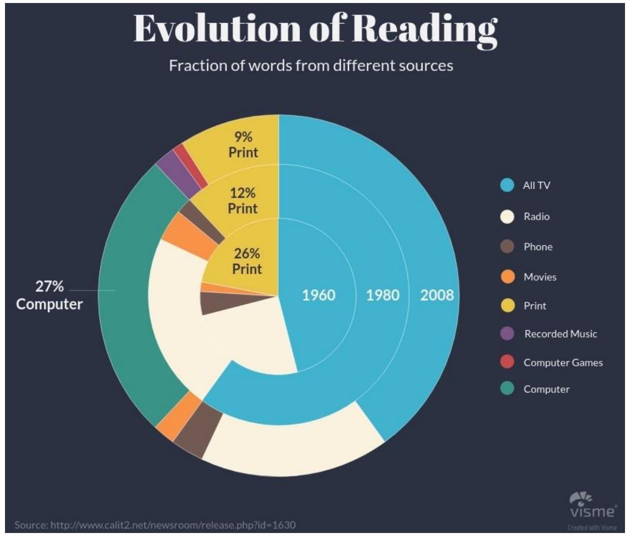


Source: https://stackoverflow.com/questions/14568819/lattice-boxplot-trellis

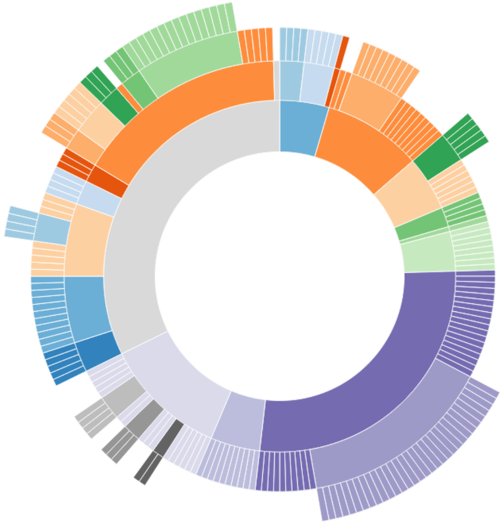


# Basic Charts

- Multi-level pie chart
  - Consists of tiers representing a separate set of data
  - Compact alternative to multiple pie charts or trellis plot

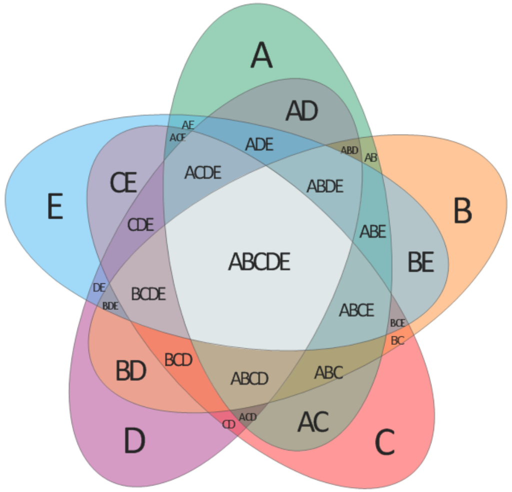


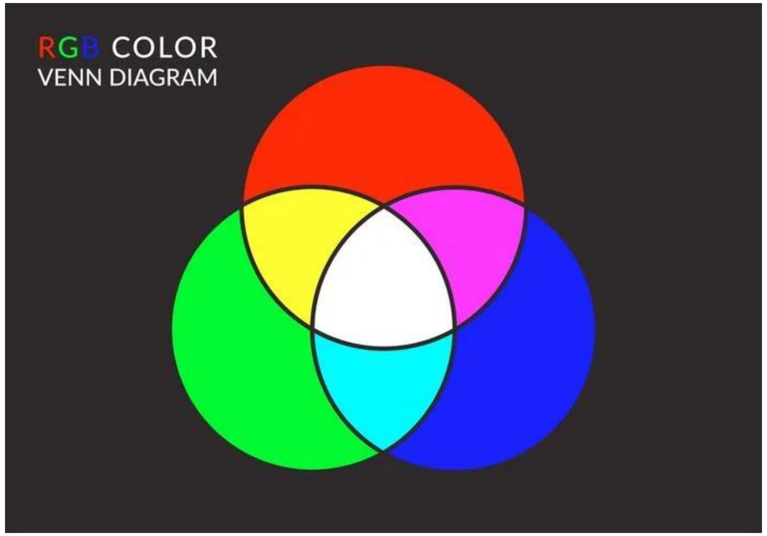
Source: http://www.calit2.net/newsroom/release.php?id=1630



# Basic Charts

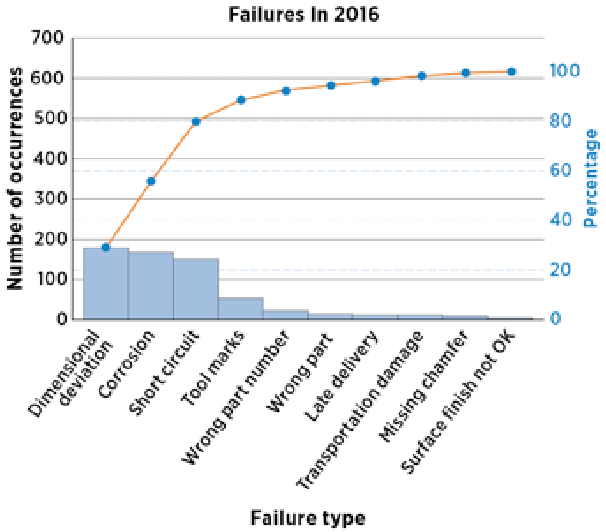
- (Euler-)Venn (logic) diagrams
  - Visualize union, intersection, difference, symmetric difference, and complement of a designated collection of sets
  - In case of scaled Venn diagrams, the relative size of each shape is proportional to the size of the group it represents

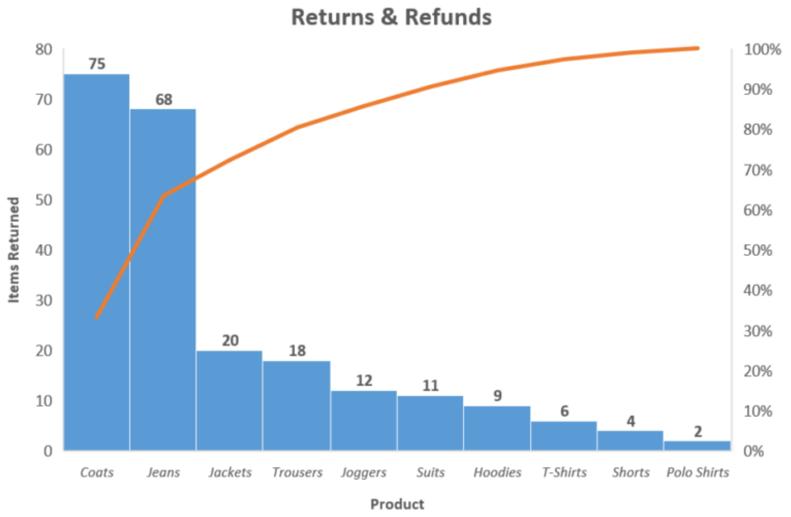




# Basic Charts

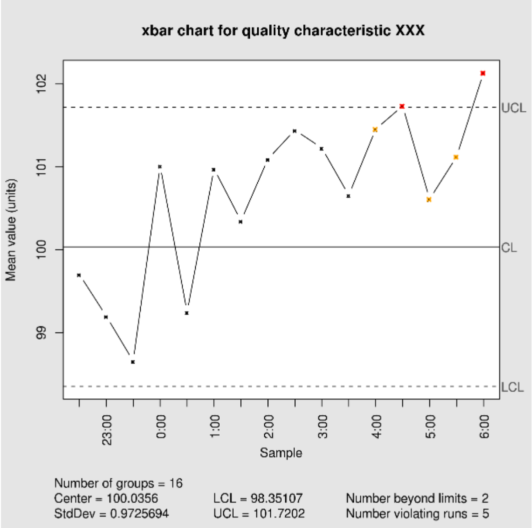
- Pareto charts
  - Combine a bar chart (individual values in descending order) and a line graph (cumulative total)
  - It highlights the most important among a typically large set of factors





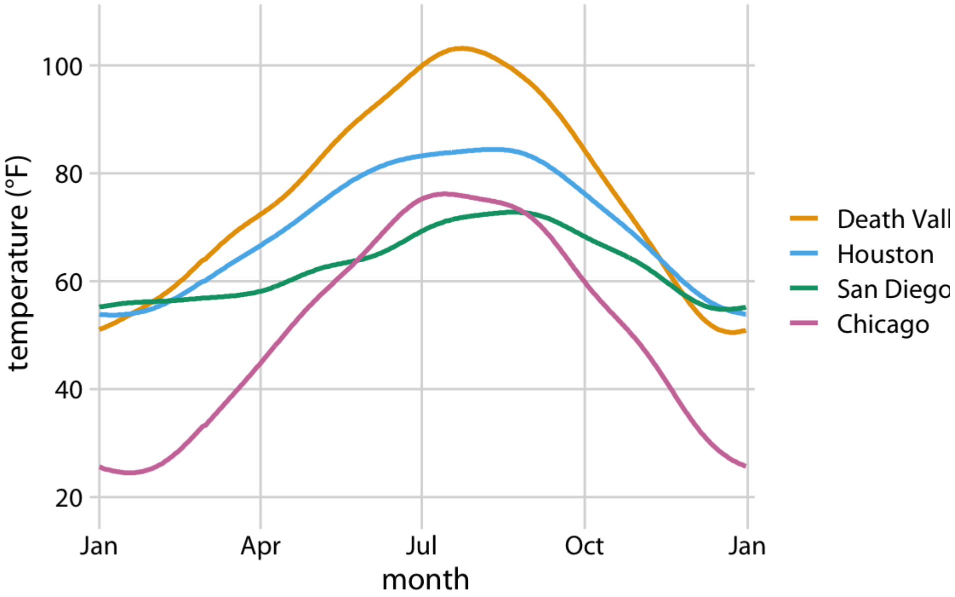
# Basic Charts

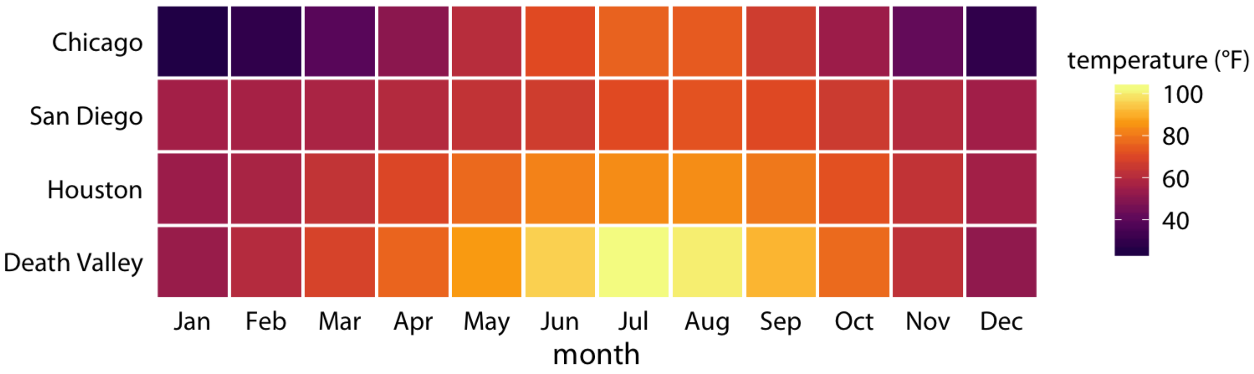
- Control (Shewhart) charts
  - Designed to monitor proces parameters with known distributions
  - CL – control line (mean or media), UCL – upper control line, LCL – lower control line (3-$\sigma$ limits, 68-95-99.7 rule)
  - Use distribution-free control charts when distribution of the underlying process is unknown



# Basic Charts

- Heat maps
  - Shows magnitude of a phenomenon as color (hue, intensity) in two dimensions
  - Uses fixed cell size for discrete phenomena and no cells for continuous

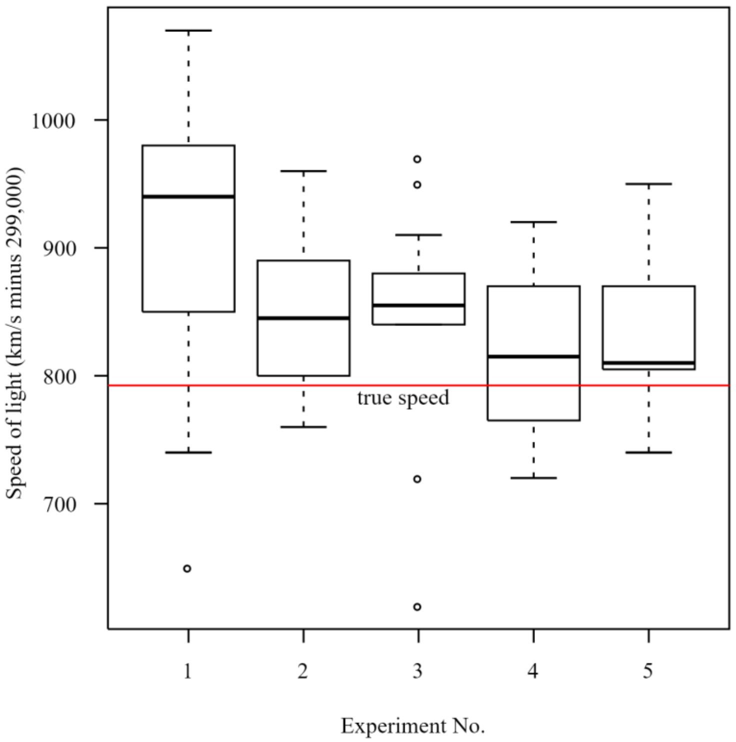




# Basic Charts

$Q_0$ … minimum
$\boldsymbol{Q_1}$ … first quartile (lower q.)
$\boldsymbol{Q_2}$ … second quartile (median)
$\boldsymbol{Q_3}$ … third quartile (upper q.)
$Q_4$ … maximum

- Box plot, box and whisker plot
  - Displays several measures of the data



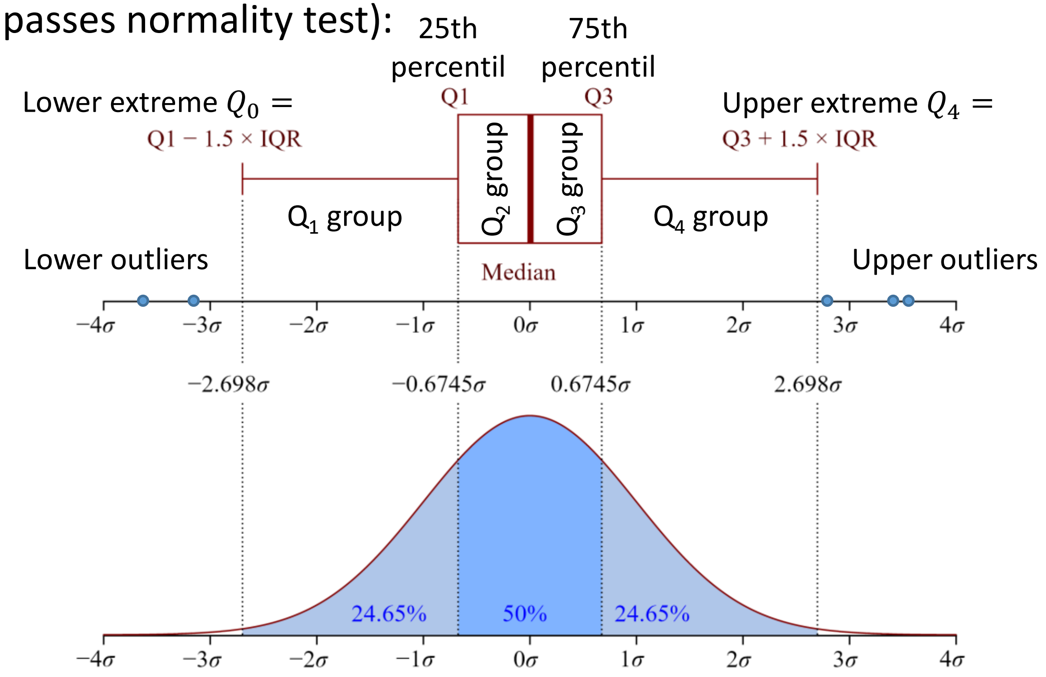
Note that if the number of observations is:
(a) *odd*, the median is the number in the middle of the list
(b) *even*, the median is the average of the middle two numbers

Data: 1, 3, 3, 4, 5, 6, 7
Q0 = 1, Q1 = 3, Q2 = 4, Q3 = 6, Q4 = 7

Data: 1, 3, 3, 4, 5, 6, 7, 8
Q0 = 1, Q1 = 3, Q2 = 4.5, Q3 = 6.5, Q4 = 8

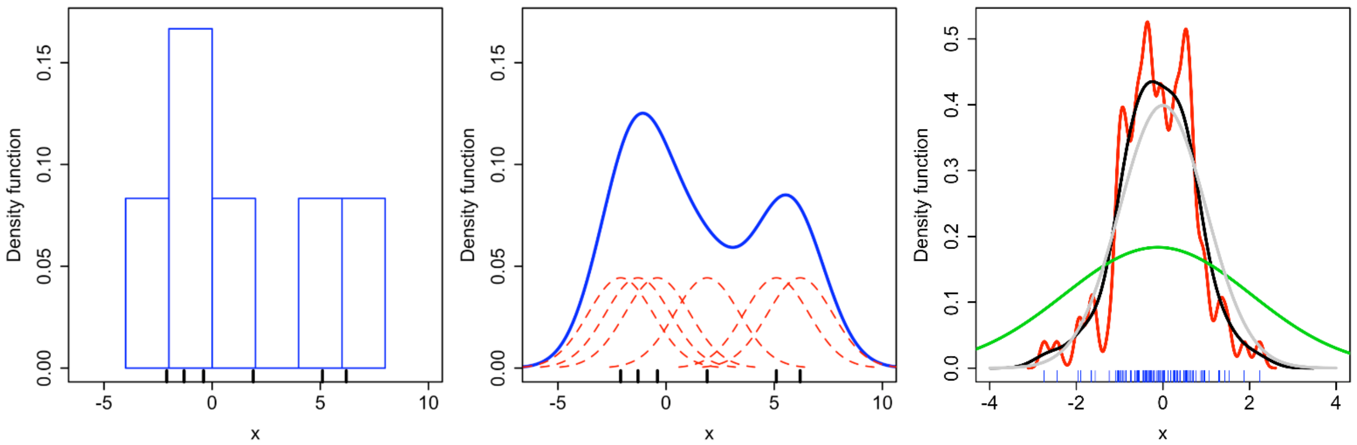
Cut the list of ordered numbers into four equal parts. Quartiles are at the cuts.

If a data set is well-modeled by a normal distribution (i.e passes normality test):



Interquartile range $\text{IQR} = Q3 - Q1 = 2 \cdot 0.6745\ \sigma$

# Basic Charts



- Kernel density plots
  - Kernel density estimate $\widehat{f_h}$ of some function $f$ is closely related to its histogram, but can be endowed with properties such as smoothness or continuity by using a suitable kernels as follows

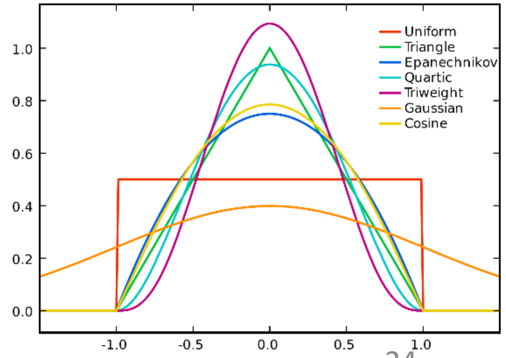
$$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right),$$

where $K$ is the kernel (non-negative function) and $h > 0$ is a smoothing parameter (bandwidth) and $K_h(x) = \frac{1}{h}K(x/h)$ is the scaled kernel

  - Kernel functions in common use are uniform, triangular, Epanechnikov (parabolic), Gaussian...

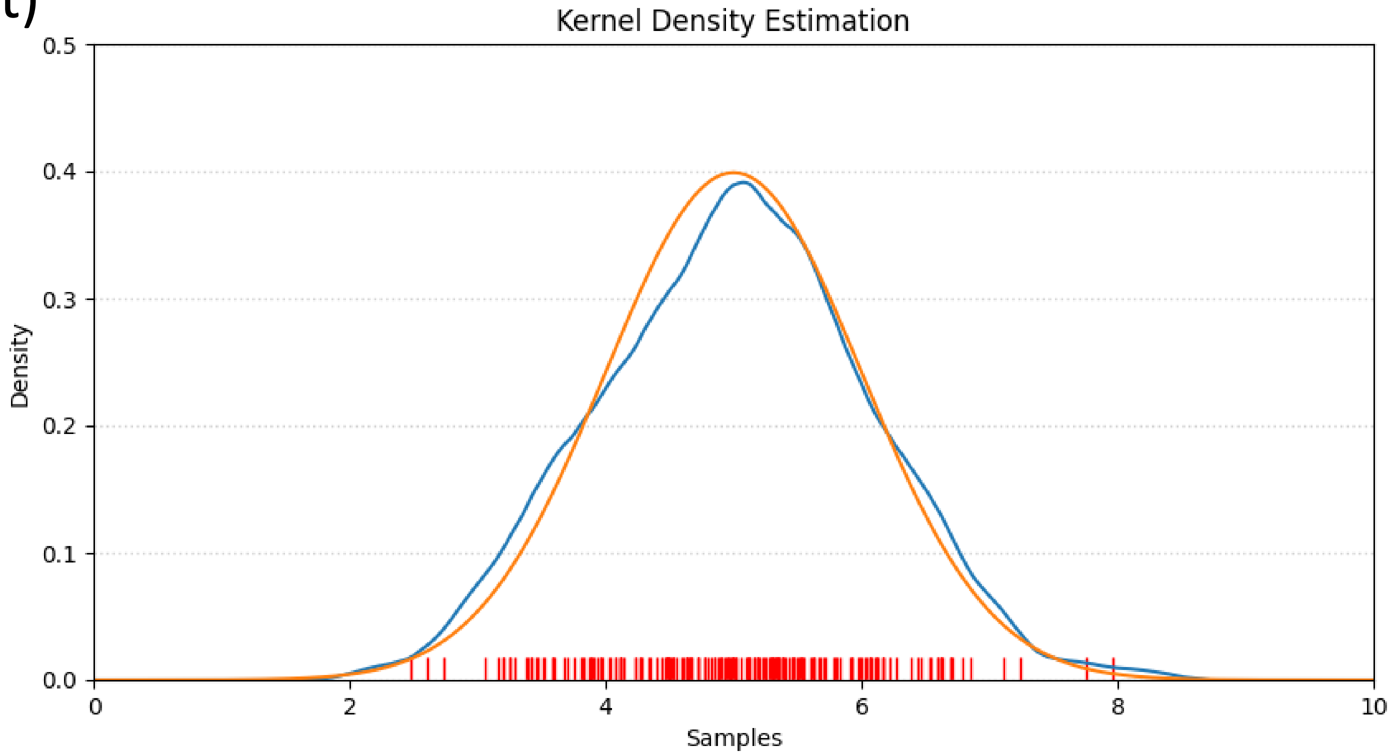
$K(u) = \frac{3}{4}(1 - u^2)$ for $|u| < 1$ otherwise 0

it holds that $\int_{-\infty}^{+\infty} K(u)\mathrm{d}u = 1$ and $K(u) = K(-u)$



# Basic Charts

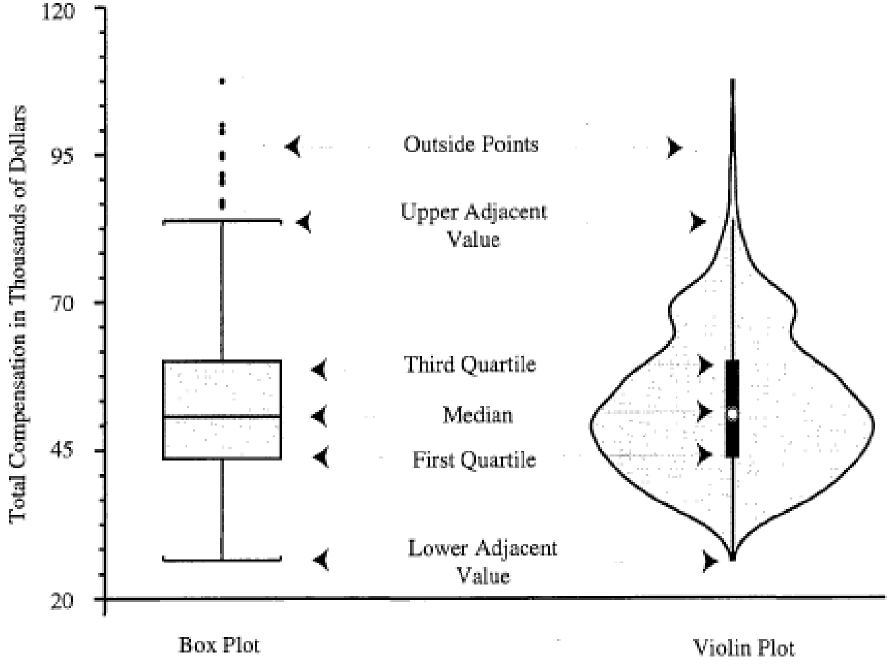
- Kernel density plots from sampled data (each red tick on x-axis represents one sample/event)



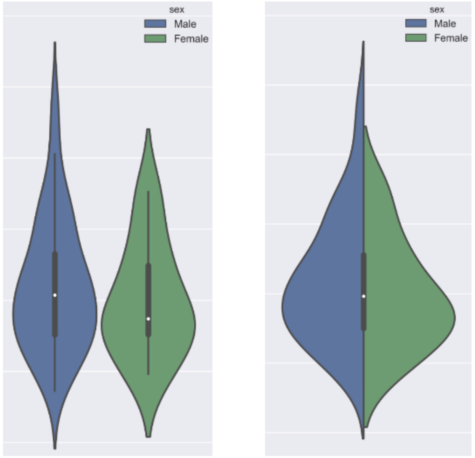
See kde.py for further reference

# Basic Charts

- Violin plot
  - Similar to box plots except that they also show the probability density of the (especially multimodal) data

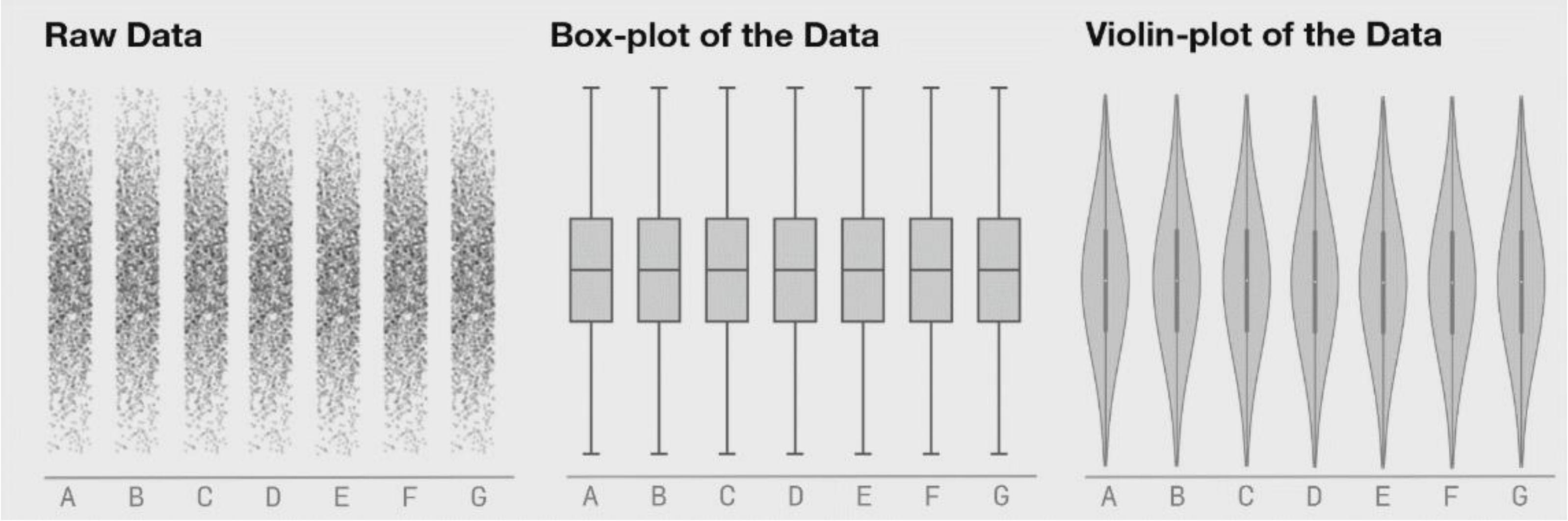


Each side of the violin may correspond to a different class



# Basic Charts

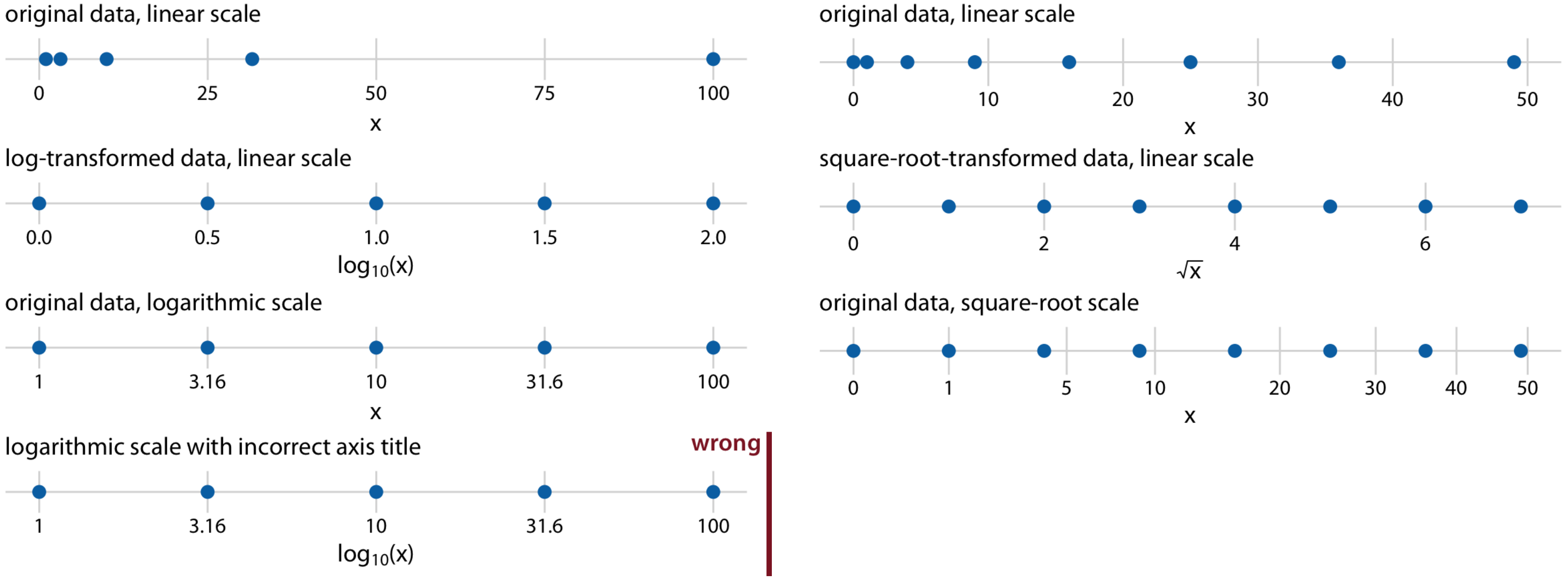
- Violin plot
  - We can modify the data in a way that the quartiles do not change, but the shape of the distribution differs dramatically



Source: https://towardsdatascience.com/violin-plots-explained-fb1d115e023d

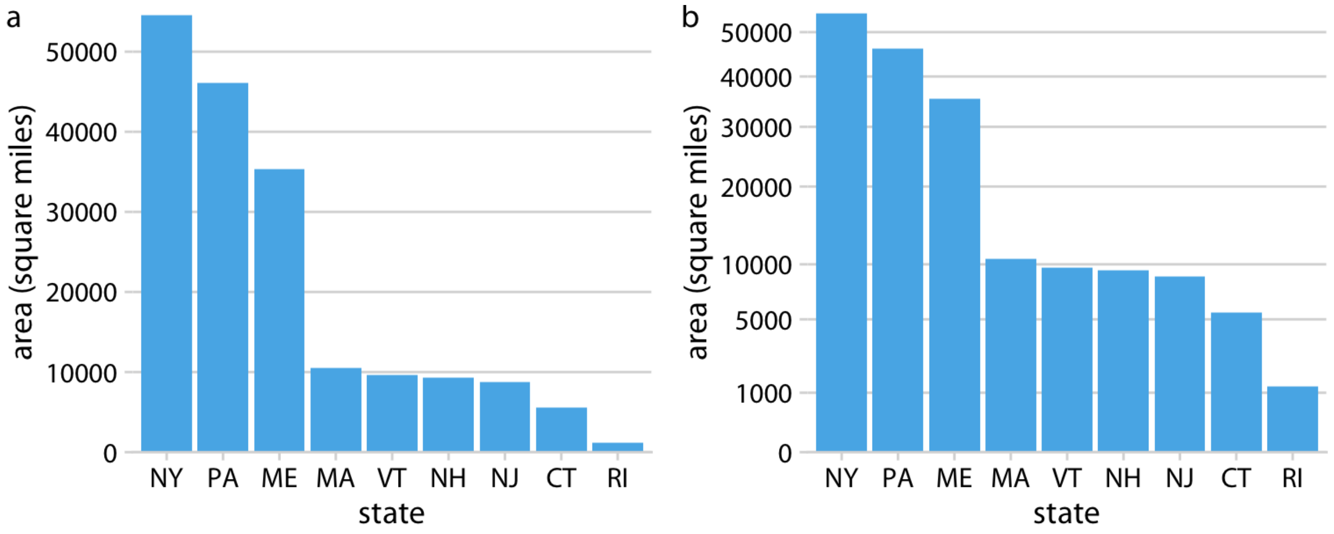
# Nonlinear Axes

- Linear scales generally provide an accurate representation of the data but there are scenarios where nonlinear scales are preferred
- The most commonly used nonlinear scale is the logarithmic scale
- The correct axis title for a log scale is the name of the variable, not the logarithm of that variable



# Nonlinear Axes

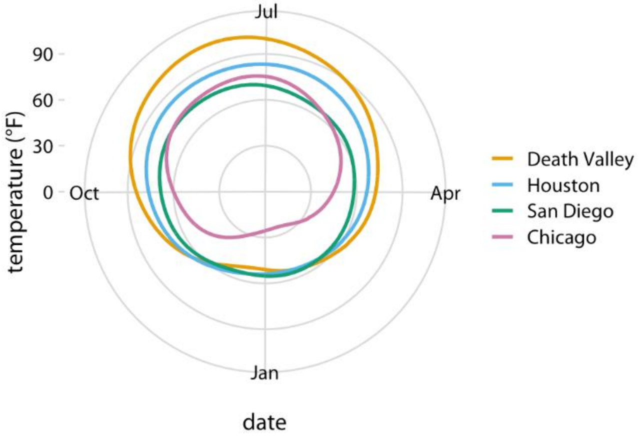
- Square-root scales have two problems
  - A unit step on  square-root scale depends on the scale value at which we are starting (a linear scale one unit step corresponds to addition or subtraction of a constant value and on a log scale it corresponds to multiplication with or division by a constant value)
  - It is unclear how to best place axis ticks on a square-root scale

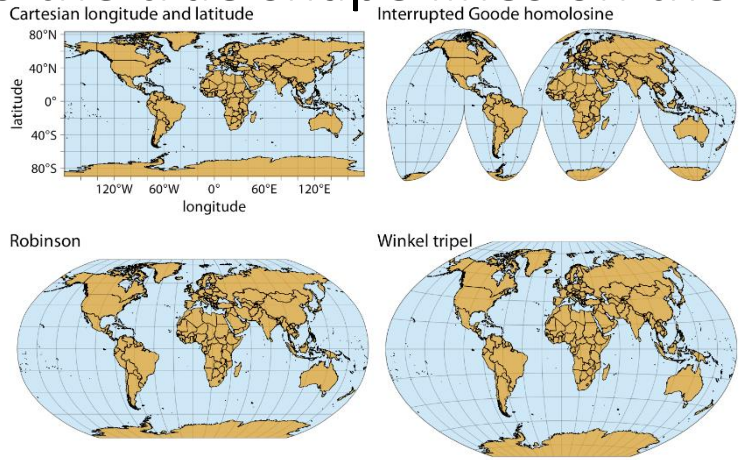


The areas of geographic regions on a square-root scale highlight the regions' linear extent from East to West or North to South. These extents could be relevant if we are wondering how long it might take to drive across a region.

# Curved Axis

- Polar coordinates
  - In the polar coordinate system, we specify positions via an angle and a radial distance from the origin
  - Polar coordinates can be useful for data of a periodic nature
- Another example are geospatial data where we use various types of non-linear projections that attempt to minimize artifacts and that strike different balances between conserving areas or angles relative to the true shape lines on the globe





# Lie Factor

- Describes the relation between the size of effect shown in a graphic and the size of effect shown in the data

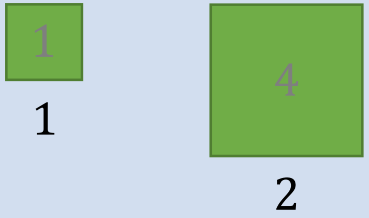
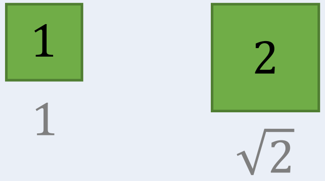
"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented."

*E. Tufte (1991)*

$$\text{lie factor} = \frac{\text{size\_of(visual effect shown in graphic)}}{\text{size\_of(actual effect shown in data)}}$$

$$\text{where size\_of} = \frac{|2^{\text{nd}}\ \text{value} - 1^{\text{st}}\ \text{value}|}{1^{\text{st}}\ \text{value}}$$

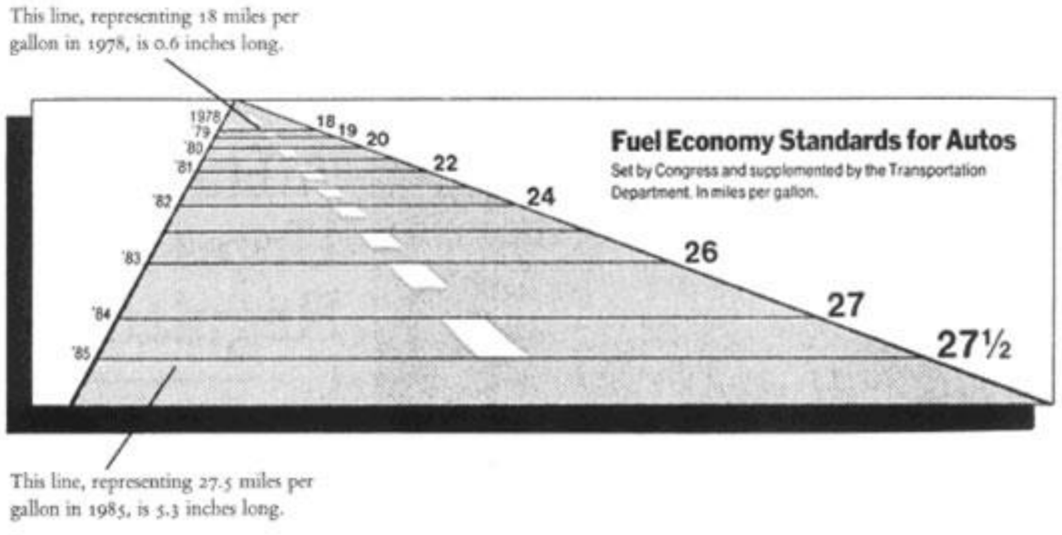
# Lie Factor

| Data Presentation | Visual Effects (Area) | Lie Factor | Result |
|---|---|---|---|
| Length of a square edge |  | $\frac{\frac{\lvert 4-1\rvert}{1}}{\frac{\lvert 2-1\rvert}{1}} = 3$ | Substantial distortion (overstating) |
| Area of a square |  | $\frac{\frac{\lvert 2-1\rvert}{1}}{\frac{\lvert 2-1\rvert}{1}} = 1$ | Integrity of a graphic and the underlying data is preserved |

The length and the area are two common options how to show one-dimensional data
Note that the area plays a role of a visual effect in this example

# Lie Factor

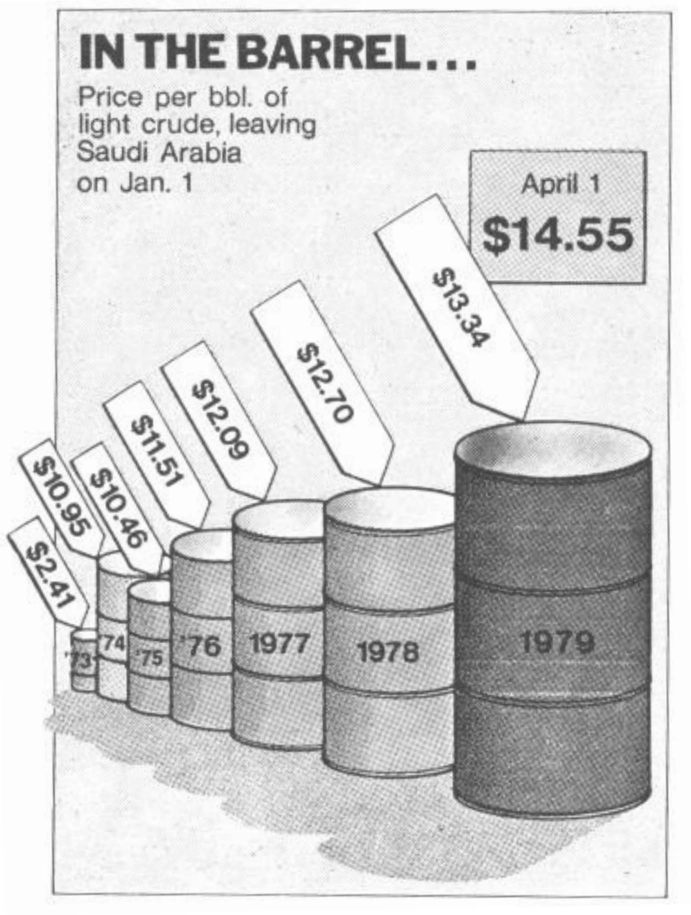
- This graphic was originally published by the NY Times. It tries to show the mandated fuel economy standards for autos set by the US Department of Transportation. The standard required an increase in mileage from 18 to 27.5, an increase of 53 %. The magnitude of increase shown in the graph is 783 %, which results in a lie factor of 14.8



Source: E. Tufte, The Visual Display of Quantitative Information, Second Edition, Graphics Press, USA, 1991.

# Lie Factor

- If we just looked at this as a 2D drawing, the lie factor would be about 9. But the metaphor presented by a 3D barrel causes the viewer to think about the volume capacity of each barrel. The capacity of the 1979 barrel is 27,000 % more than the 1973 barrel, even though the price only increased by 554 % during that time – a Lie Factor of 27,000 / 554 = 48.8



Source: E. Tufte, The Visual Display of Quantitative Information, Second Edition, Graphics Press, USA, 1991.

# Data-Ink Ratio

- The data-ink ratio is the proportion of ink that is used to present actual data compared to the total amount of ink (or pixels) used in the entire graphic

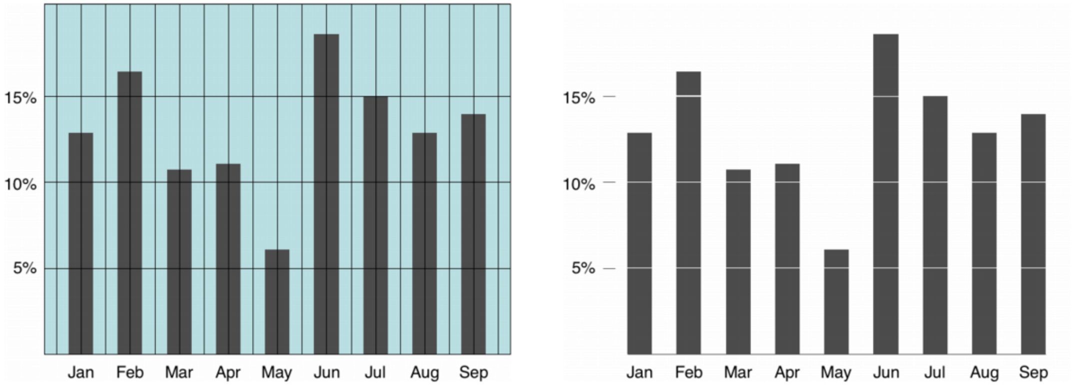
$$\text{data}-\text{ink ratio} = \frac{\text{data}-\text{ink}}{\text{total ink used to print graphic}}$$

- Good graphics should include only data-ink
- Non-data-ink may act as a distraction to the viewers
- Several things can contribute to distraction in data visualization: use of 3D effects, background images, shadow effects, unnecessary borders, and unnecessary grid lines

# Data-Ink Ratio

- E. Tufte provides five laws to data-ink:
    1. Above all else show the data
    2. Maximize the data-ink ratio
    3. Erase non-data-ink
    4. Erase redundant data-ink
    5. Revise and edit
- In other words, simplify charts and graphs to the point where they are clear and understandable but not a step further

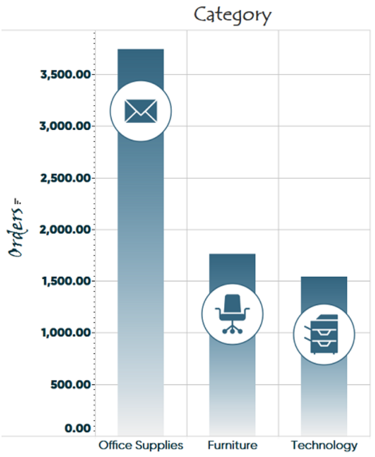
Chartjunk vs Tufte's minimalist design of bar-graphs



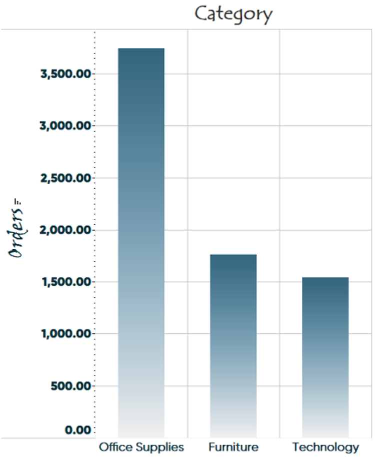
Source: https://infovis-wiki.net/wiki/Data-Ink_Ratio

# Data-Ink Ratio

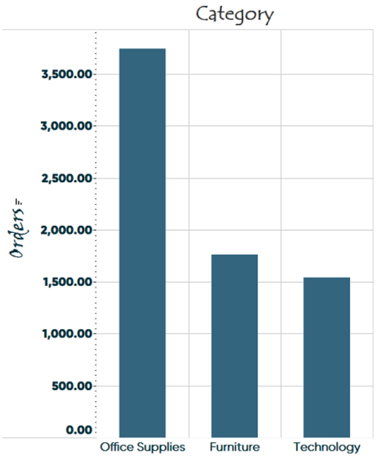
1. remove chart junk (except the graphics improving memorability or with additional value)



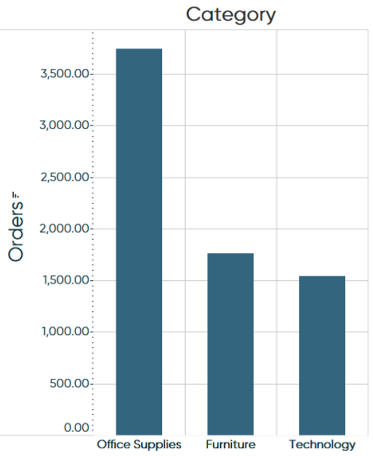
2. remove effects (gradients)



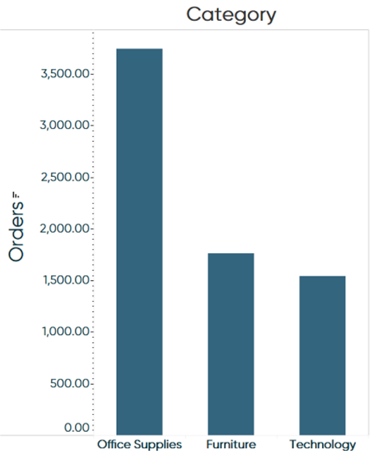
3. remove varying font styles and formats (bolds)



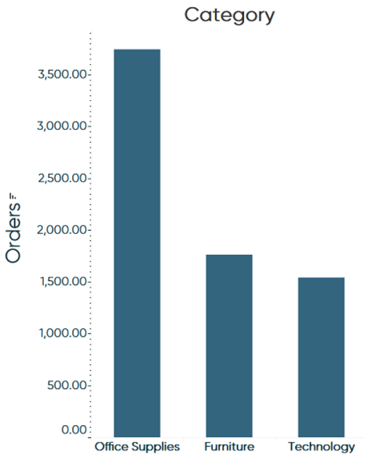
4. remove the grid



5. remove the borders



6. is this really ok?

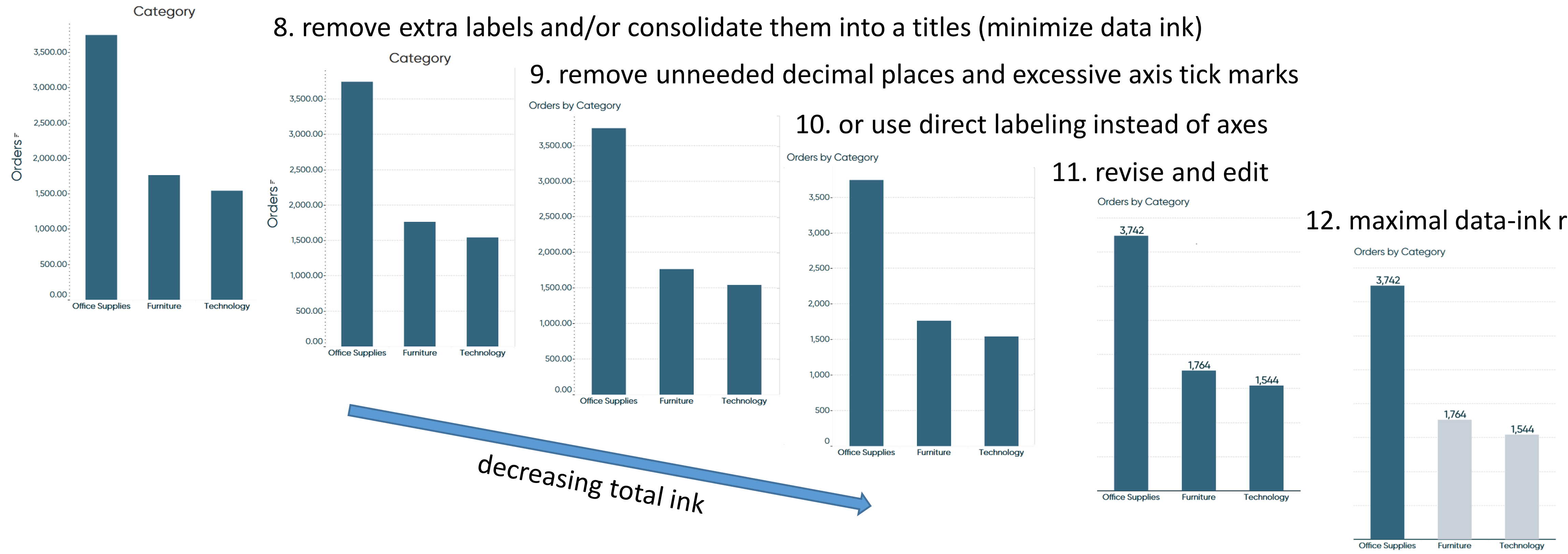


decreasing total ink

Source: https://playfairdata.com/data-ink-ratio-animation-and-how-to-apply-it-in-tableau/

# Data-Ink Ratio

7. add a light dotted grid to ease comparison



8. remove extra labels and/or consolidate them into a titles (minimize data ink)

9. remove unneeded decimal places and excessive axis tick marks

10. or use direct labeling instead of axes

11. revise and edit

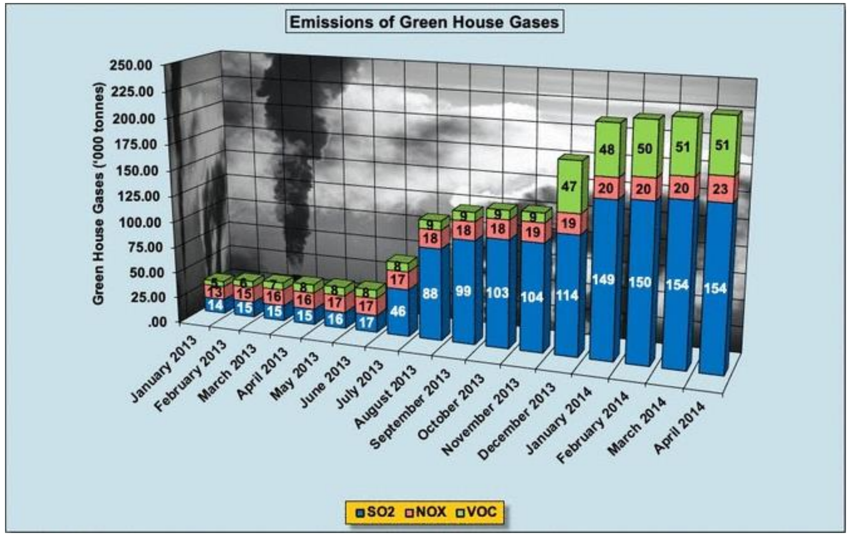
12. maximal data-ink ratio

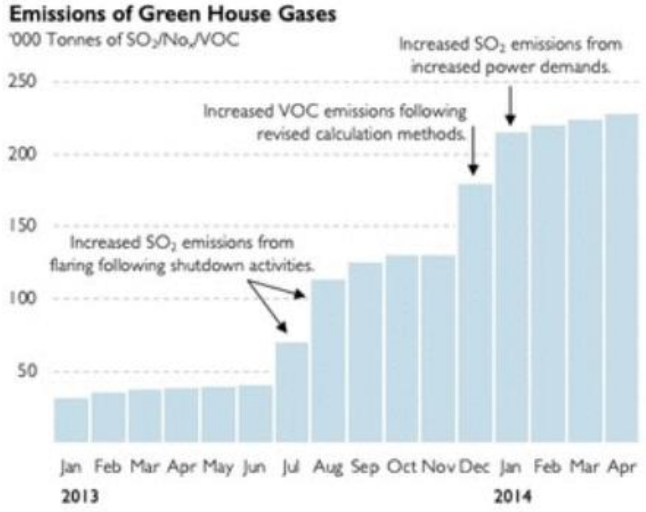
decreasing total ink

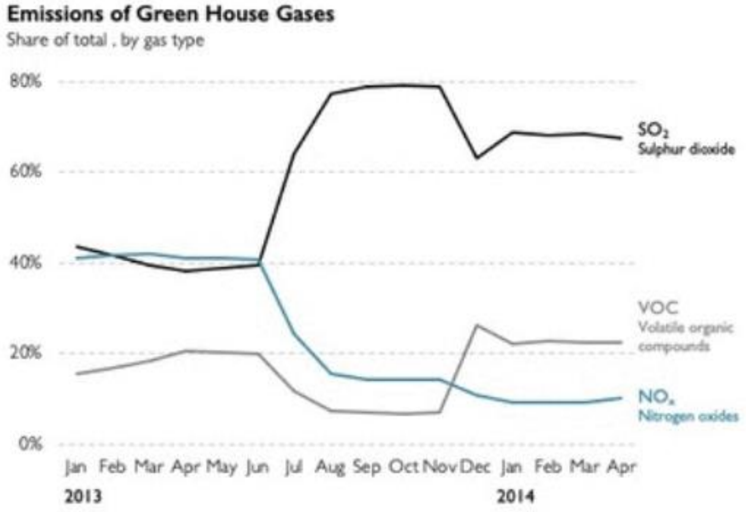
Source: https://playfairdata.com/data-ink-ratio-animation-and-how-to-apply-it-in-tableau/

# Data-Ink Ratio Examples

- Low vs high data-ink ratio – use words, numbers, and drawing together







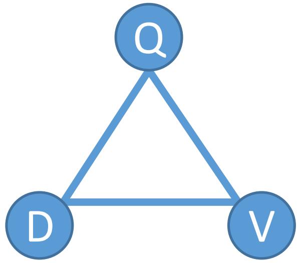
Source: https://simplexct.com/data-ink-ratio

# Chartjunk

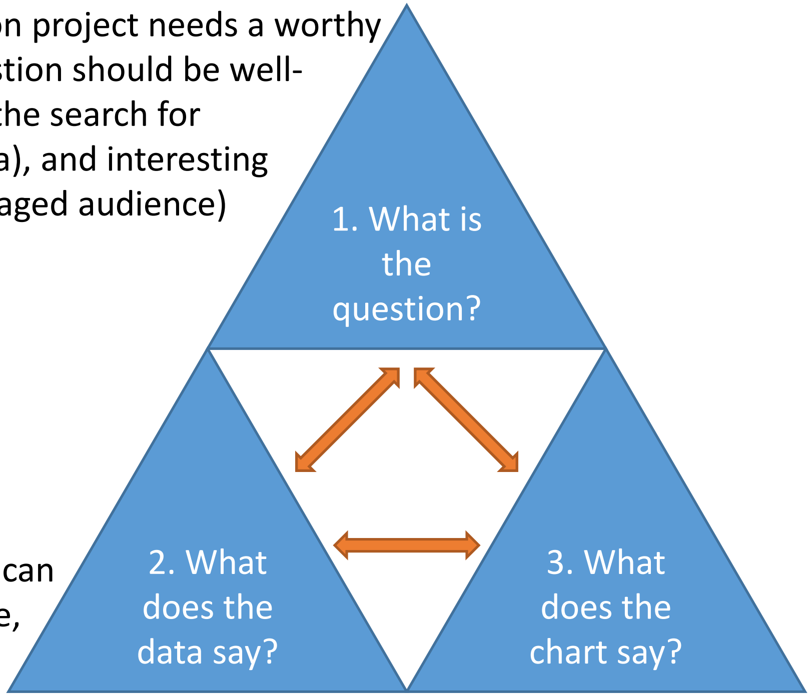
- Chartjunk is a term for unnecessary elements or decorations that distract the viewers from the communicated information
  - Unnecessary expansion of 2D information into a 3D visualization
  - Distractive textures and gradients instead of solid colors
  - Using illustrations to represent data
  - Cluttered backgrounds
  - Using a complex color maps/schemes to convey data
  - Attempting to put two or more charts into one
  - Including non-essential information into a chart
  - Disregarding data visualization conventions

# Trifecta Checkup Framework

- There are 8 possible states of chart critiques



Data visualization project needs a worthy cause. The Question should be well-posed (focuses the search for appropriate data), and interesting (ensures an engaged audience)



1. What is the question?

2. What does the data say?

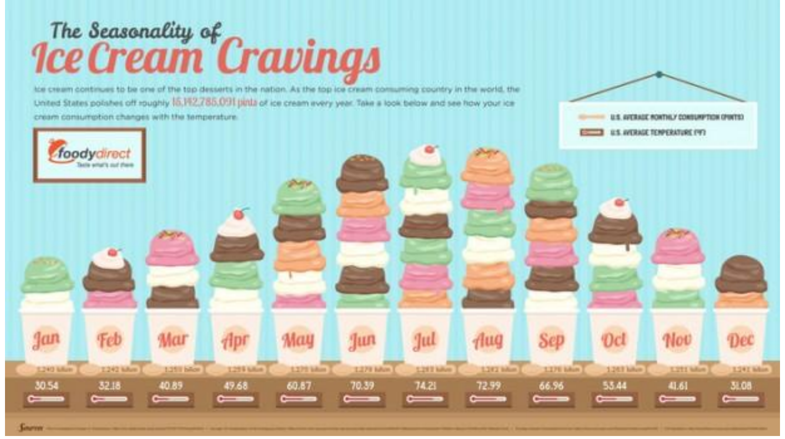
3. What does the chart say?

The Data should be relevant to the Question being addressed. Relevance can often be augmented by reducing noise, removing errors or transformations.

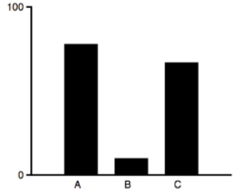
The Visual elements (chart) should represent the Data in a clear, concise manner, addressing the Question directly.

Source: https://junkcharts.typepad.com/junk_charts/junk-charts-trifecta-checkup-the-definitive-guide.html

# Effects of Visual Embellishment

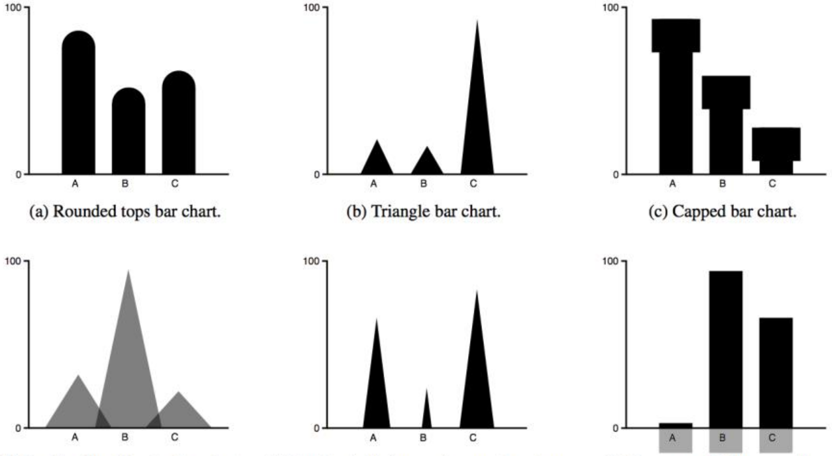


http://blog.visual.ly/exploring-perception-bar-charts/



(g) Baseline bar chart.

vs.



(a) Rounded tops bar chart.

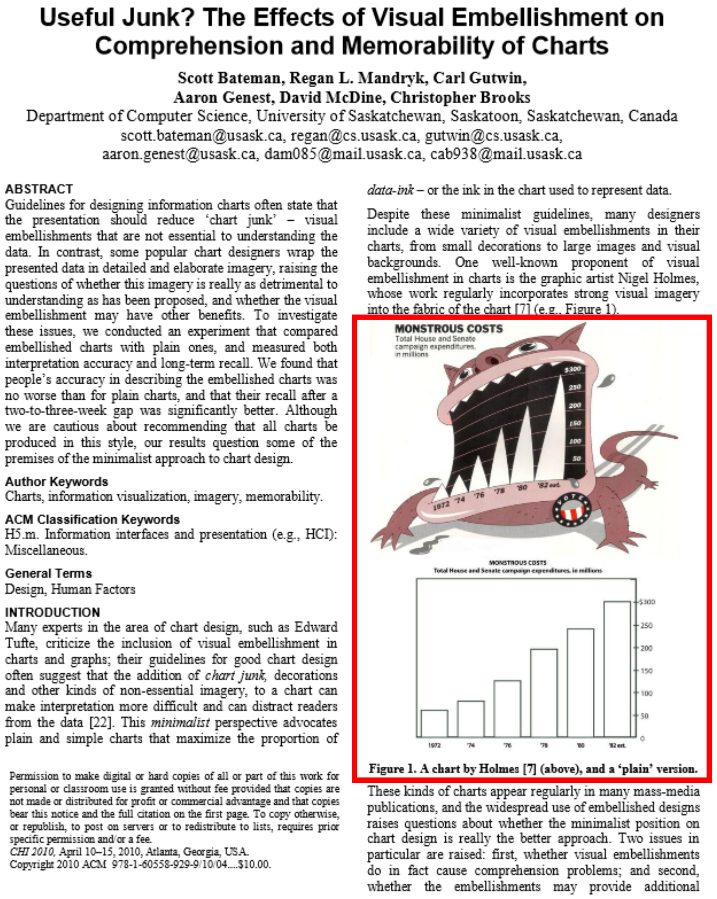
(b) Triangle bar chart.

(c) Capped bar chart.

(d) Overlapping triangles bar chart.

(e) Quadratically increasing area bar chart.

(f) Bars extend below zero line.



### Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts

**Scott Bateman, Regan L. Mandryk, Carl Gutwin,**
**Aaron Genest, David McDine, Christopher Brooks**
Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
scott.bateman@usask.ca, regan@cs.usask.ca, gutwin@cs.usask.ca,
aaron.genest@usask.ca, dam085@mail.usask.ca, cab938@mail.usask.ca

**ABSTRACT**

Guidelines for designing information charts often state that the presentation should reduce 'chart junk' – visual embellishments that are not essential to understanding the data. In contrast, some popular chart designers wrap the presented data in detailed and elaborate imagery, raising the questions of whether this imagery is really as detrimental to understanding as has been proposed, and whether the visual embellishment may have other benefits. To investigate these issues, we conducted an experiment that compared embellished charts with plain ones, and measured both interpretation accuracy and long-term recall. We found that people's accuracy in describing the embellished charts was no worse than for plain charts, and that their recall after a two-to-three-week gap was significantly better. Although we are cautious about recommending that all charts be produced in this style, our results question some of the premises of the minimalist approach to chart design.

**Author Keywords**
Charts, information visualization, imagery, memorability.

**ACM Classification Keywords**
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

**General Terms**
Design, Human Factors

**INTRODUCTION**

Many experts in the area of chart design, such as Edward Tufte, criticize the inclusion of visual embellishment in charts and graphs; their guidelines for good chart design often suggest that the addition of *chart junk*, decorations and other kinds of non-essential imagery, to a chart can make interpretation more difficult and can distract readers from the data [22]. This *minimalist* perspective advocates plain and simple charts that maximize the proportion of *data-ink* – or the ink in the chart used to represent data.

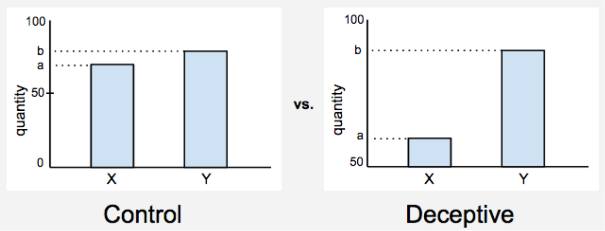
Despite these minimalist guidelines, many designers include a wide variety of visual embellishments in their charts, from small decorations to large images and visual backgrounds. One well-known proponent of visual embellishment in charts is the graphic artist Nigel Holmes, whose work regularly incorporates strong visual imagery into the fabric of the chart [7] (e.g., Figure 1).

**Figure 1. A chart by Holmes [7] (above), and a 'plain' version.**

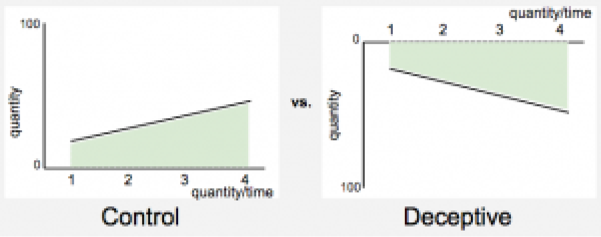
These kinds of charts appear regularly in many mass-media publications, and the widespread use of embellished designs raises questions about whether the minimalist position on chart design is really the better approach. Two issues in particular are raised: first, whether visual embellishments do in fact cause comprehension problems; and second, whether the embellishments may provide additional

*CHI 2010*, April 10–15, 2010, Atlanta, Georgia, USA.
Copyright 2010 ACM 978-1-60558-929-9/10/04....$10.00.

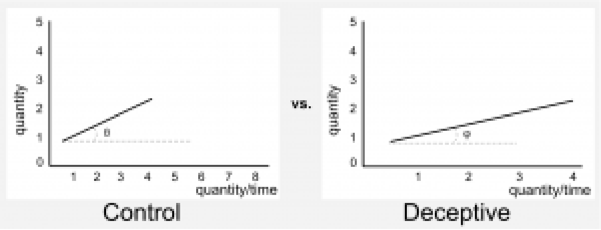
# Effects of Visual Embellishment



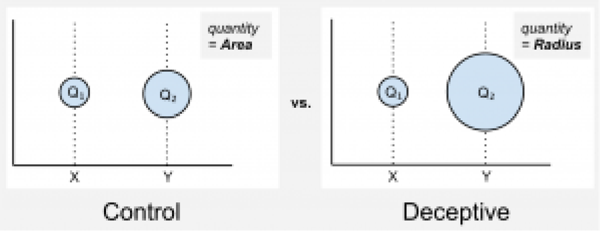
Truncated axis



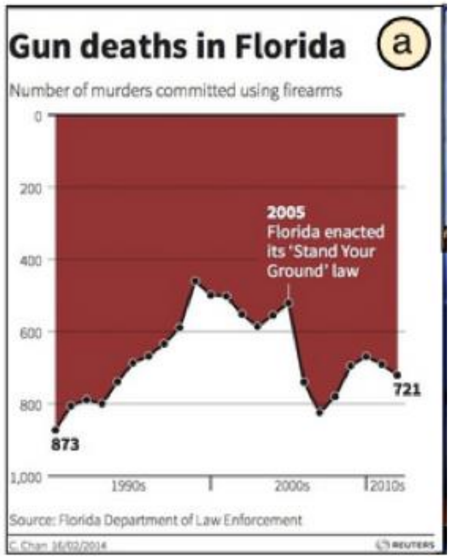
Invertex axis



Aspect ratio



Area mapping



http://fellinlovewithdata.com/research/deceptive-visualizations