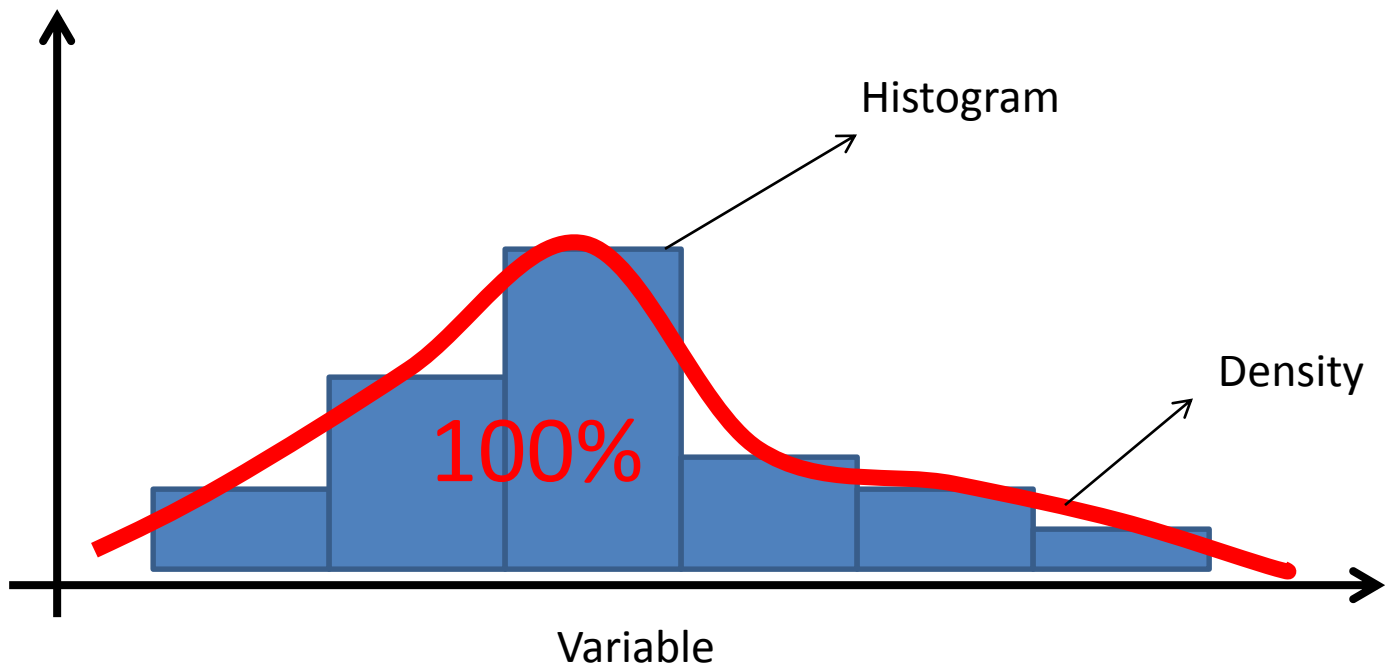# Distribution shapes

Galton board

Recomended video and applet:

http://www.youtube.com/watch?v=6YDHBFVIvIs
http://www.disfrutalasmatematicas.com/datos/quincunce.html
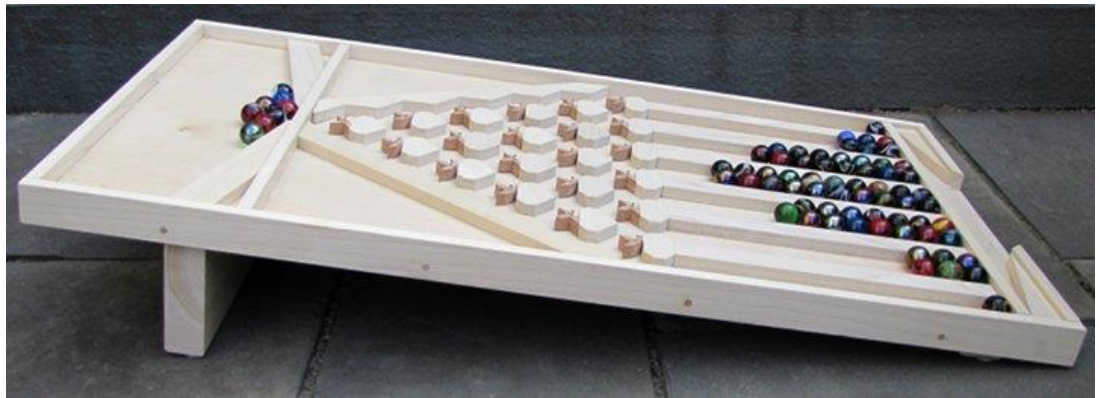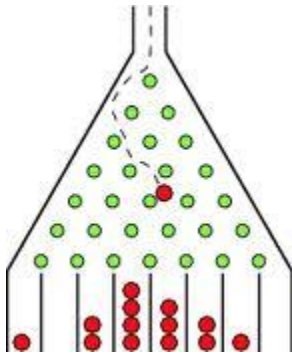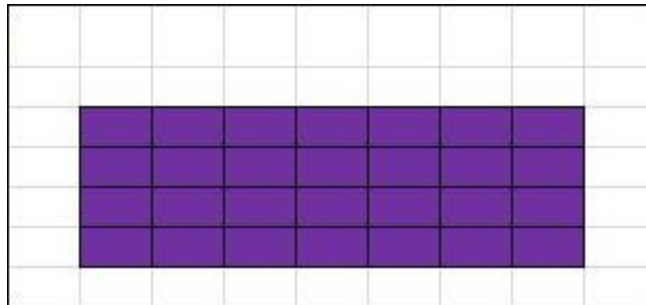
Distribution shapes

A frequency distribution is **symmetric** when a vertical line can be drawn through the middle of a graph of the distribution and the resulting halves are approximately the mirror images.
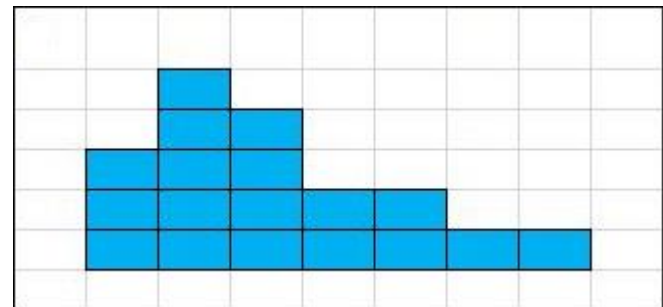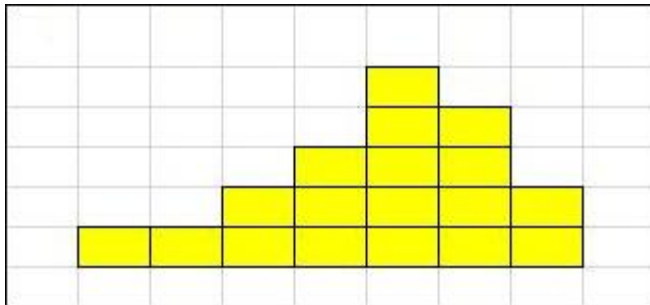
Distribution shapes

A frequency distribution is **uniform** (or **rectangular**) when all entries, or classes, in the distribution have equal frequencies. A uniform distribution is also symmetric.
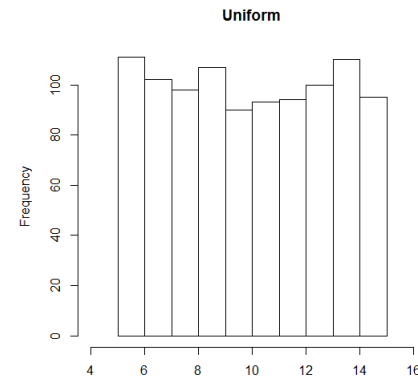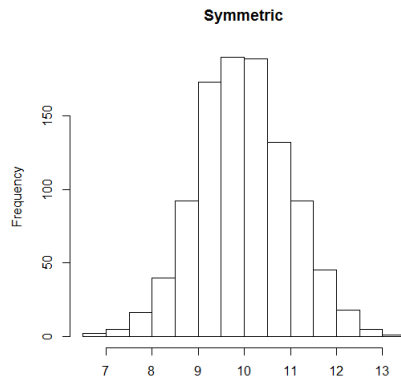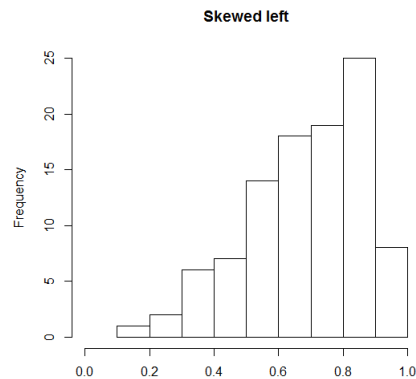
Distribution shapes

A frequency distribution is skewed if the "tail" of the graph elongates more to one side than to the other. A distribution is **skewed left** (**negatively skewed**) if its tail extends to the left. A distribution is **skewed right** (**positively skewed**) if its tail extends to the right.
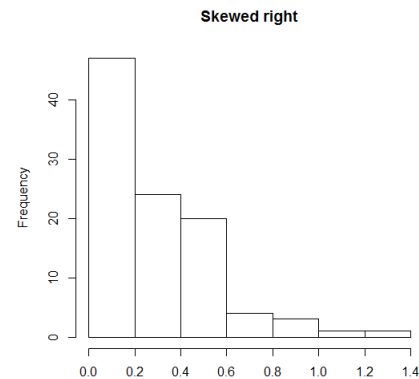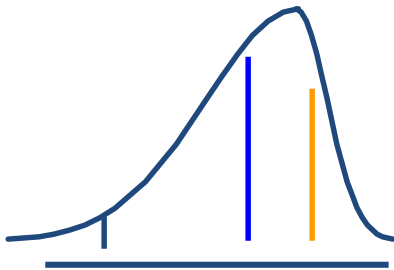
Summary of distribution shapes

Symmetric

Uniform

Mean ≈ Median

Skewed left

Skewed right

Mean < Median < Mode

Mode < Median < Mean

Distribution Shape and Boxplot
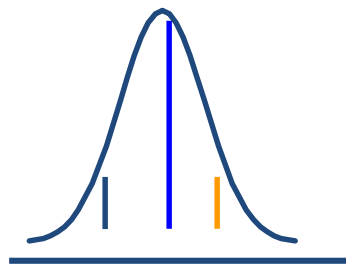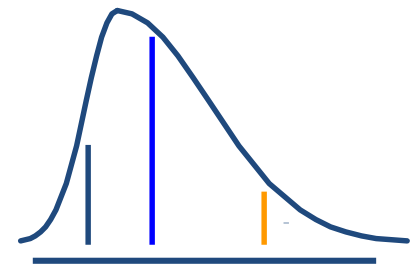
Pearson Correlation Coefficient r

Measures the strength of the linear relationship between two quantitative variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Features of Correlation Coefficient

- Unit free

- Ranges between –1 and 1

- The closer to –1, the stronger the negative linear relationship

- The closer to 1, the stronger the positive linear relationship

- The closer to 0, the weaker any positive linear relationship

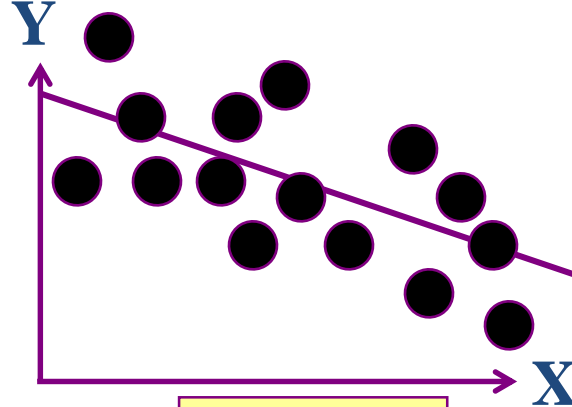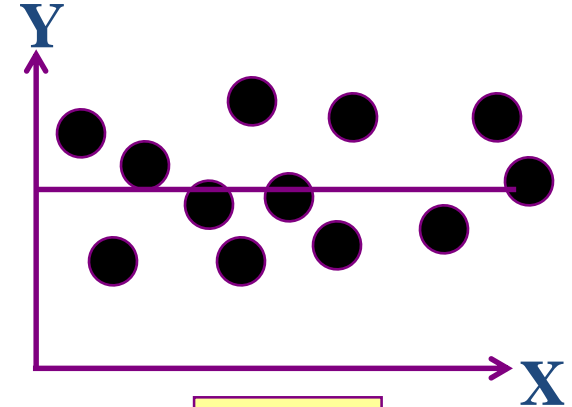# Correlation coefficient

Shows strength and direction of correlation

Strong ← Weak | Weak → Strong

-1.0   0.5   0   0.5   1.0

Negative correlation   Positive correlation
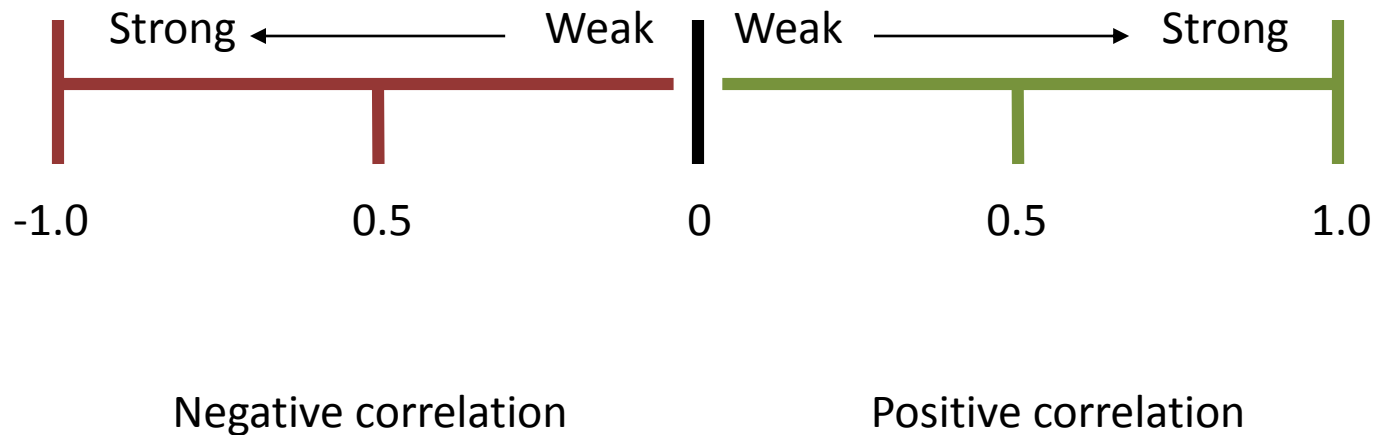
Scatter Plots of Data with Various Correlation Coefficients
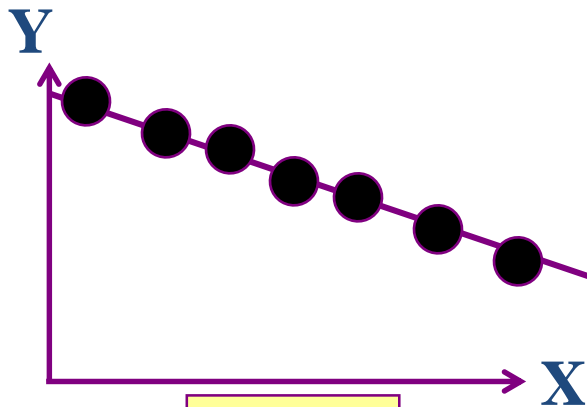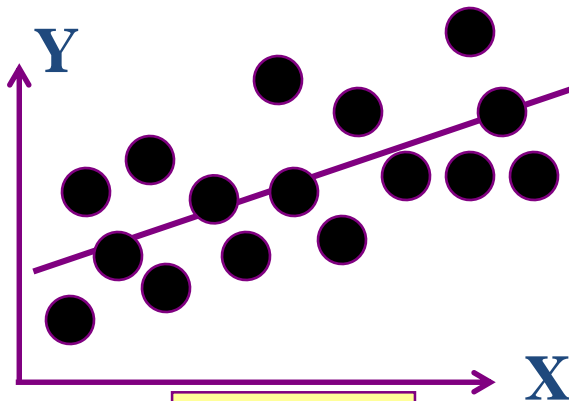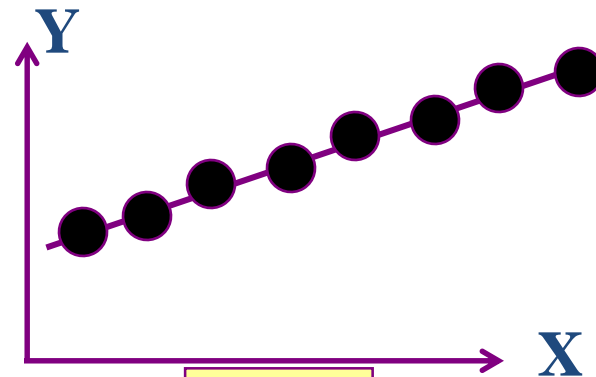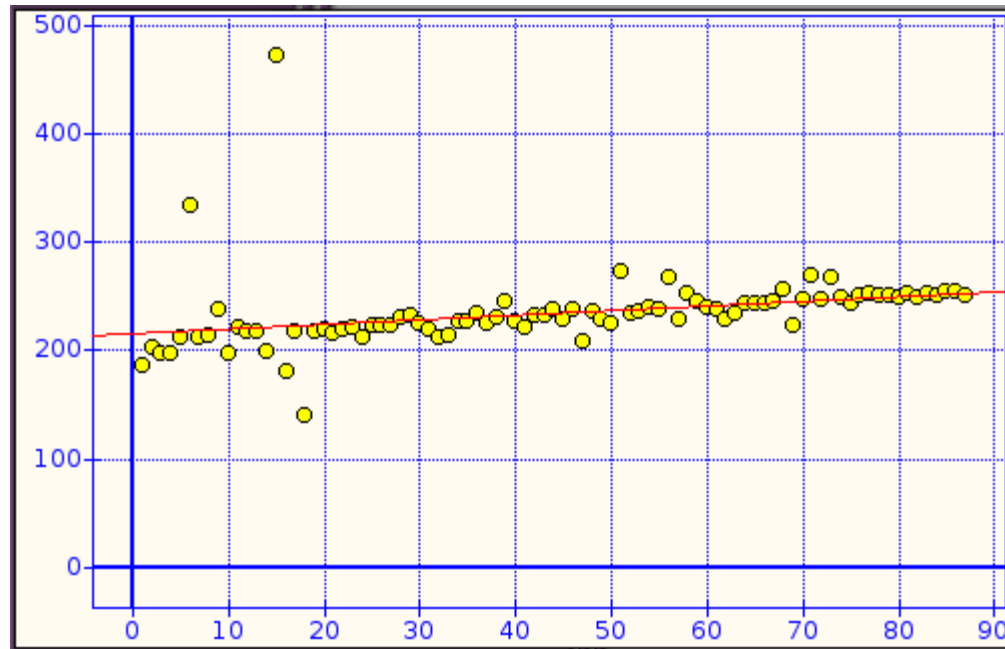
Coeficiente de correlación de Spearman ρ

El coeficiente de correlación de Spearman, ρ (rho) es una medida de la correlación entre dos variables aleatorias x e y continuas.

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

donde D es la diferencia entre los correspondientes estadísticos de orden de $x - y$. N es el número de parejas.
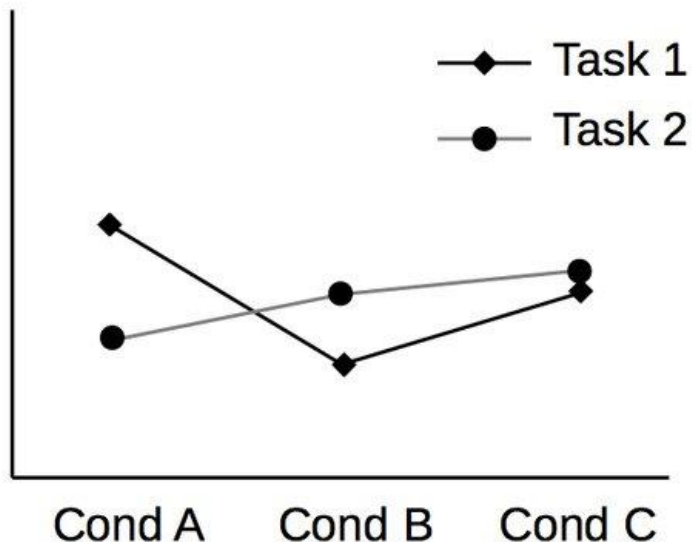
El coeficiente de correlación de Spearman es menos sensible que el de Pearson para los valores muy lejos de lo esperado. En este ejemplo: Pearson = 0.30706 Spearman = 0.76270

Tips

What your scientific figure looks like, vs. what the audience sees

Chartjunk

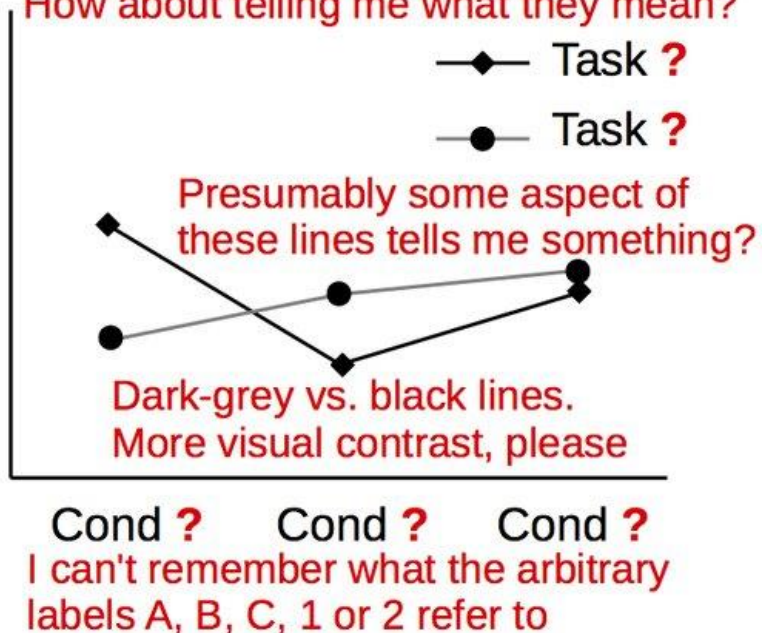**Chartjunk** refers to all visual elements in charts and graphs that are not necessary to comprehend the information represented on the graph, or that distract the viewer from this information.

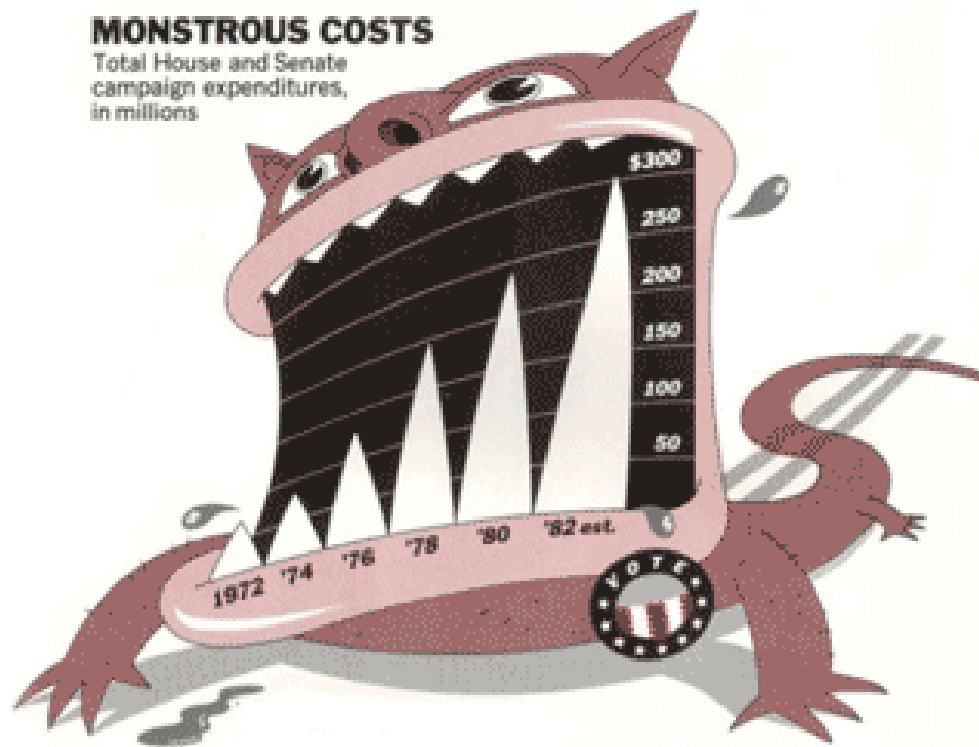https://en.wikipedia.org/wiki/Chartjunk
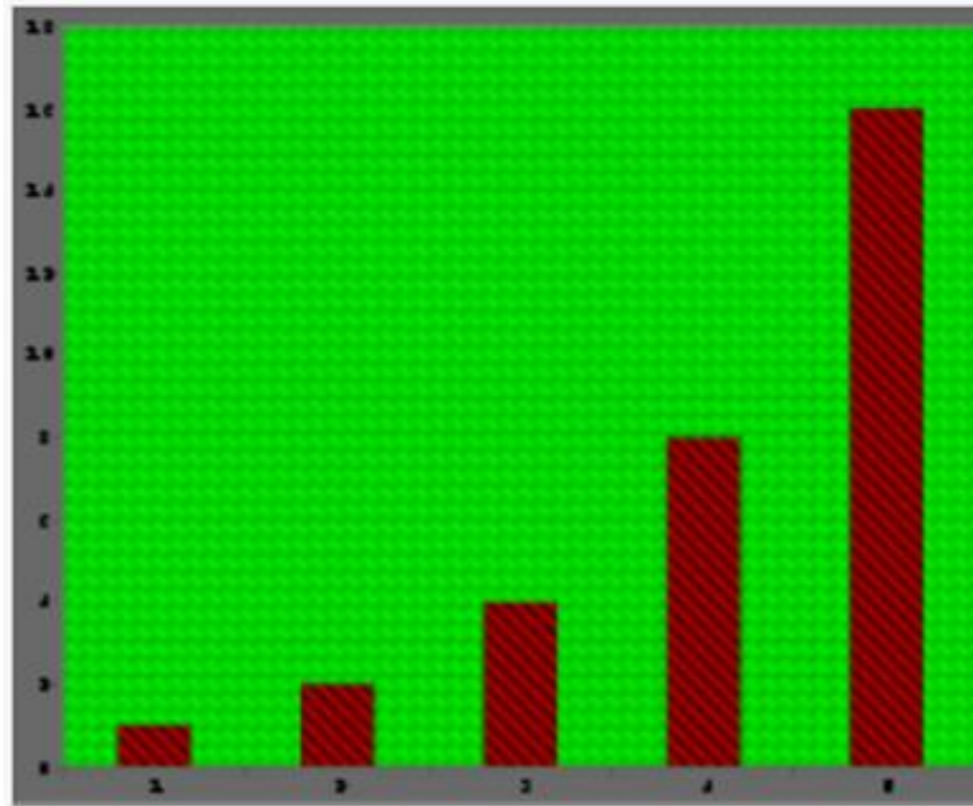
An example of a chart containing gratuitous chartjunk. This chart uses a large area and a lot of "ink" (many symbols and lines) to show only five hard-to-read numbers, 1, 2, 4, 8, and 16.

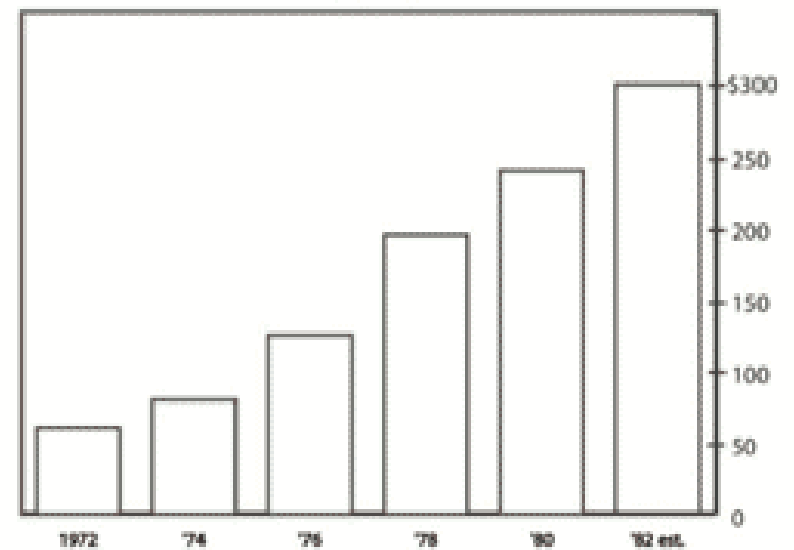A map with chartjunk: the gradients inside each province
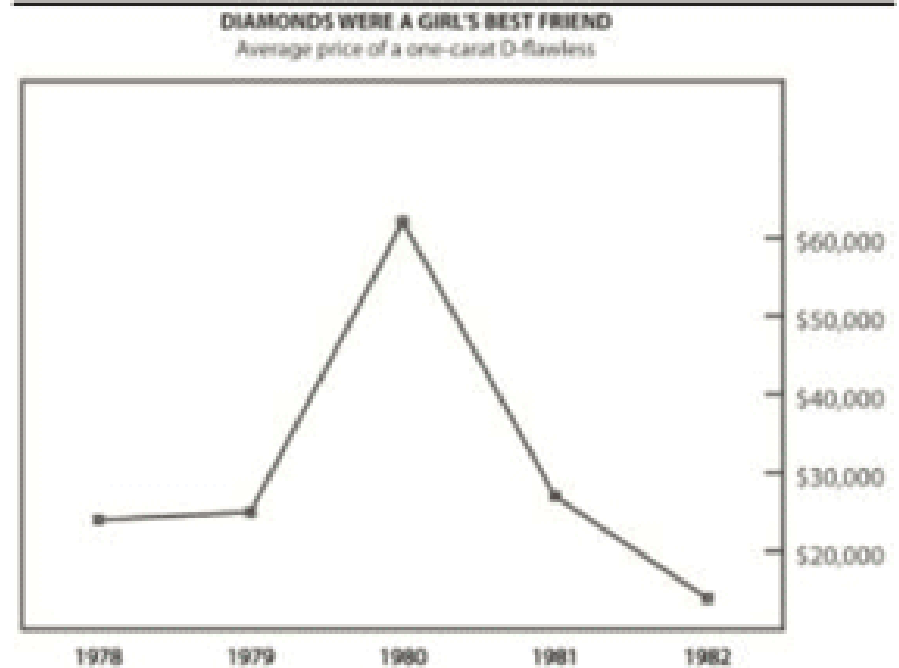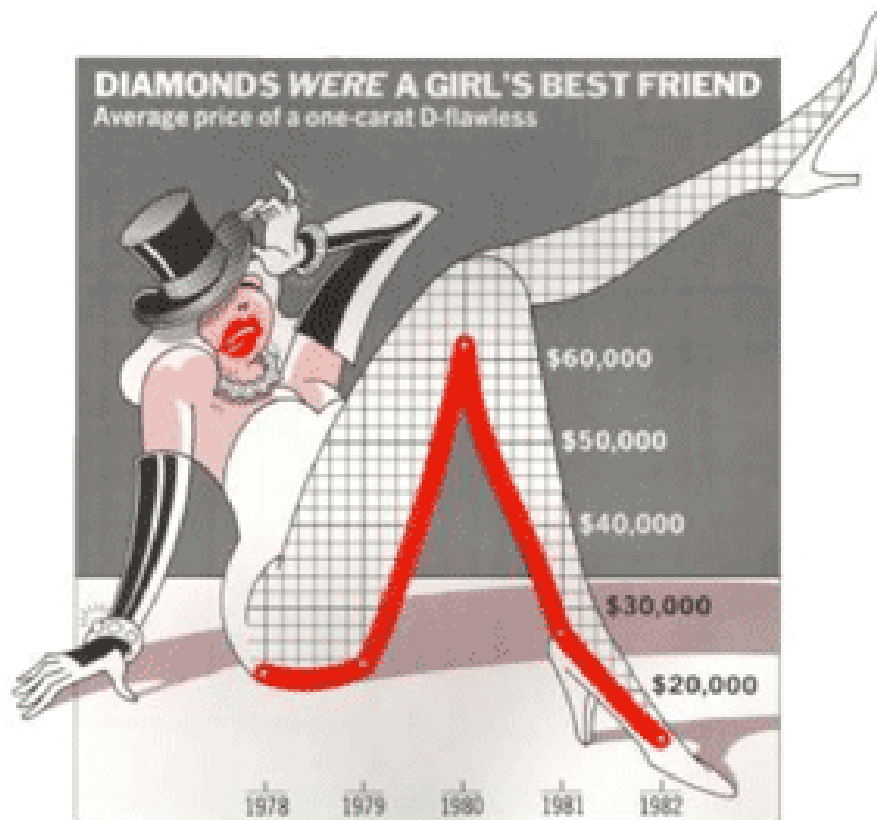
MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions

MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions

DIAMONDS *WERE* A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless

$60,000
$50,000
$40,000
$30,000
$20,000

1978 1979 1980 1981 1982

DIAMONDS WERE A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless

$60,000
$50,000
$40,000
$30,000
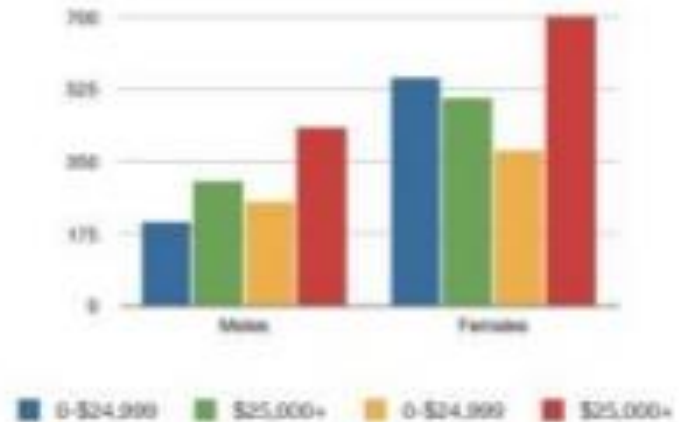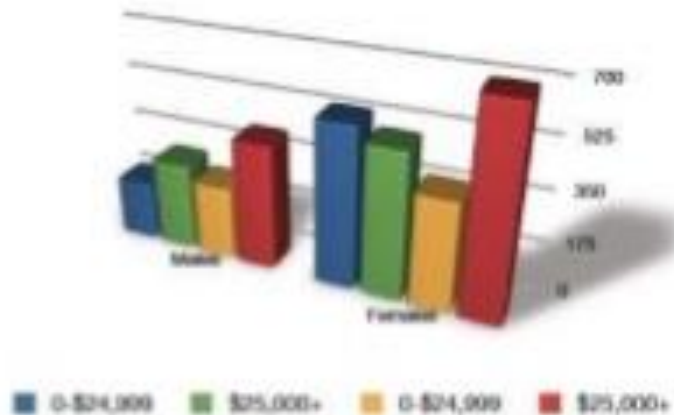$20,000

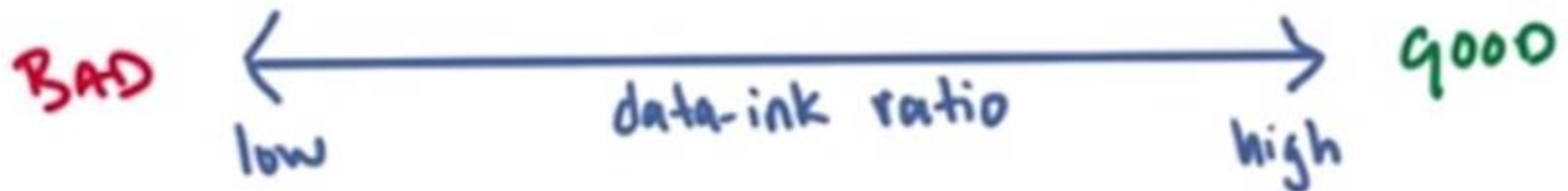1978 1979 1980 1981 1982

Data-ink ratio formula

Maximize Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

Data-ink ratio formula

$$\text{data-ink ratio} = \frac{\text{ink used to describe the data}}{\text{ink used to describe everything else}}$$

BAD $\longleftrightarrow$ good

low   data-ink ratio   high

The value of *telling stories* with data as opposed to merely displaying it.

What do you think about the picture?