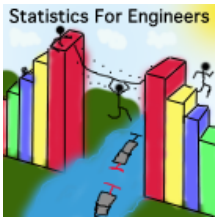
Statistics for Engineers



© Kevin Dunn, 2015
kevin.dunn@mcmaster.ca
http://learnche.mcmaster.ca/

Univariate Data Analysis

Copyright, sharing, and attribution notice

We appreciate:

- if you let us know about **any errors** in the slides
- **any suggestions to improve the notes**

All of the above can be done by writing to

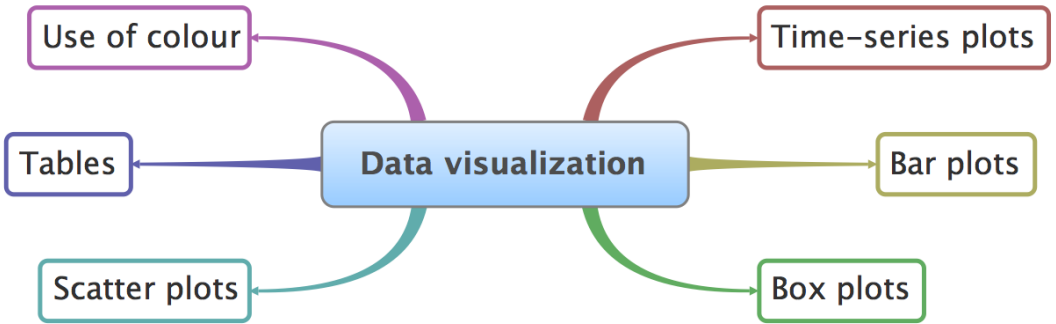
kevin.dunn@mcmaster.ca

or anonymous messages can be sent to Kevin Dunn at

http://learnche.mcmaster.ca/feedback-questions

If reporting errors/updates, please quote the current revision number:

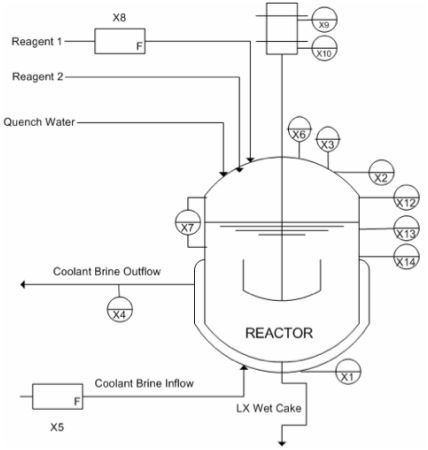
Topics covered in this section on data visualization



Some ways you can impress your colleagues and manager in the future

- *Co-worker*: Here are the yields from a batch system for the last 3 years (1256 data points), can you help me:
  - understand more about the time-trends in the past 3 year?
  - efficiently summarize the yield from all batches run in 2010?
- *Manager*: effectively summarize the (a) number and (b) types of defects on 17 aluminum grades for the past 12 months
- *Yourself*: 24 different variables being measured *vs* time (5 readings per minute, over 300 minutes) for each batch we produce; how can we visualize these 36,000 data points?
  - see next slides

Batch systems generate large quantities of *extremely* valuable data

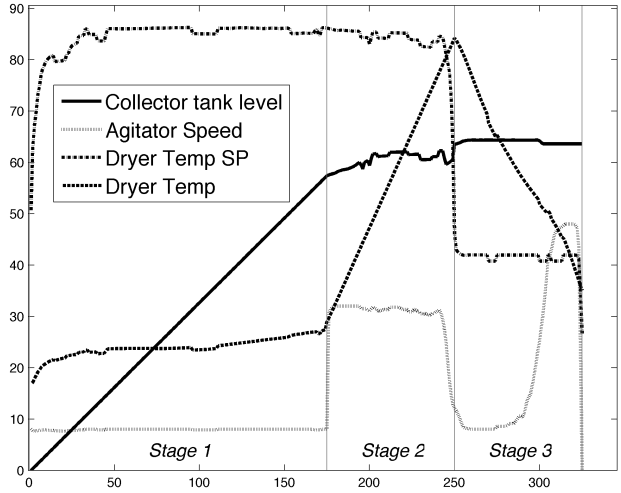


[From Cecilia Rodrigues' M.A.Sc thesis, 2006, McMaster University, used with permission]

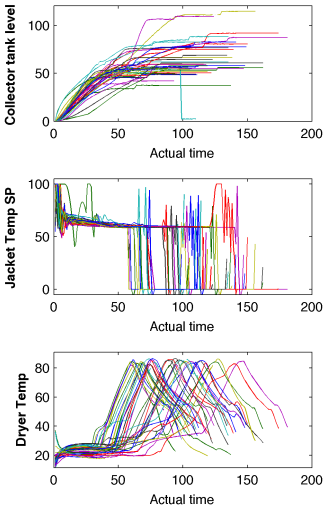
[Flickr: #2516220152]

Batch systems generate large quantities of *extremely* valuable data

Data from a single batch



Data from many batches



Recommended references for this section: Data visualization

1. Edward Tufte, *Envisioning Information*, Graphics Press, 1990. (10th printing in 2005)
2. Edward Tufte, *The Visual Display of Quantitative Information*, Graphics Press, 2001.
3. Edward Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative*, 2nd edition, Graphics Press, 1997.
4. William Cleveland, *Visualizing Data*, and *The Elements of Graphing Data*, Hobart Press; 2nd edition, 1994.
5. Stephen Few, *Show Me the Numbers*, and "Now You See It", Analytics Press.
6. Su, It's easy to produce chartjunk using Microsoft Excel 2007 but hard to make good graphs, *Computational Statistics and Data Analysis*, **52** (10), 4594-4601, 2008, http://dx.doi.org/10.1016/j.csda.2008.03.007

Why bother learning about this topic: it's too easy!

This class might seem too easy; too obvious. It is!

- The human eye and brain are excellent at pattern recognition, sorting through signal and noise.
- We can easily cope with bad plots; but good plots save time and show a clearer, more honest picture.
- Cliches: "Let the data speak for themselves", "Plot the data"

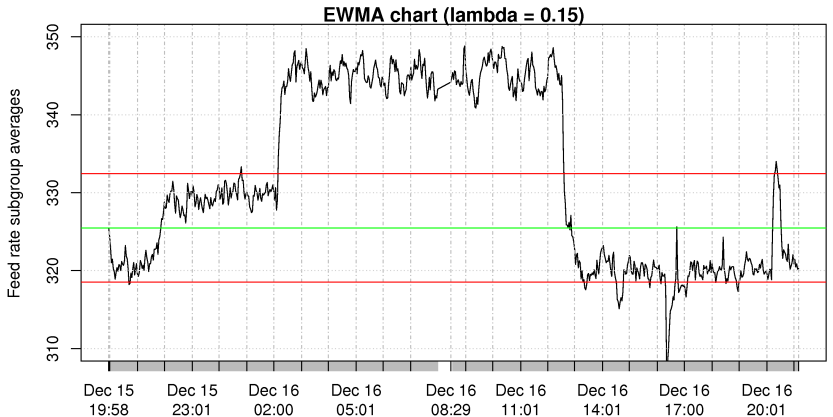
Strong suggestion: find a bad plot (journal publications, an old lab report that you have written); upload it to the forums and criticize the plot. Why is it bad?

We need good plots to make decisions quickly, correctly, and confidently

[Flickr: #7848902570]

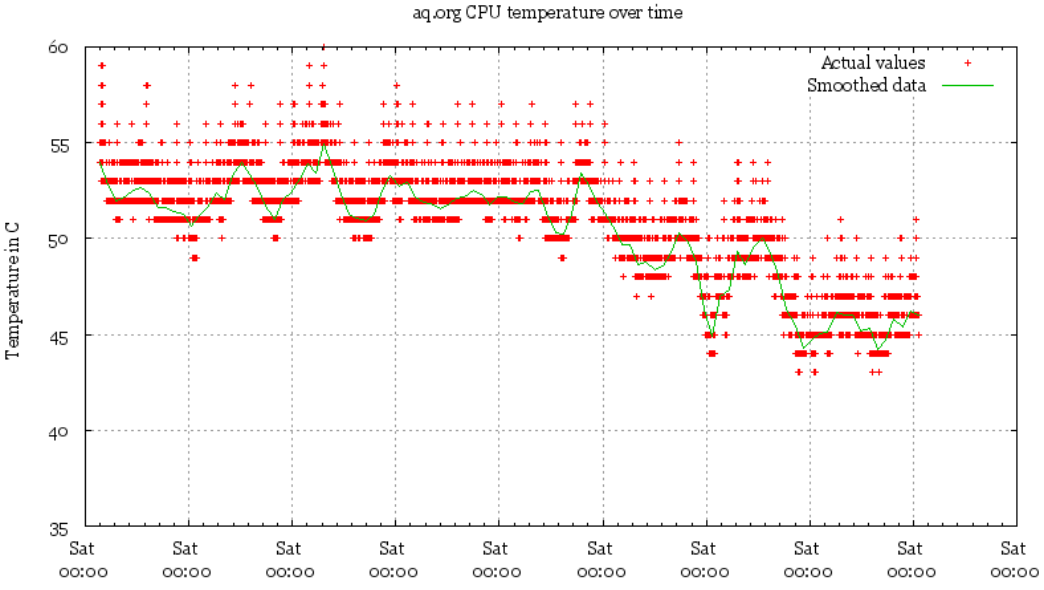
Time-series plots show a univariate piece of information in 2 dimensions

- (usually) have the horizontal x-axis show time or sequence order
- the other axis: the data values

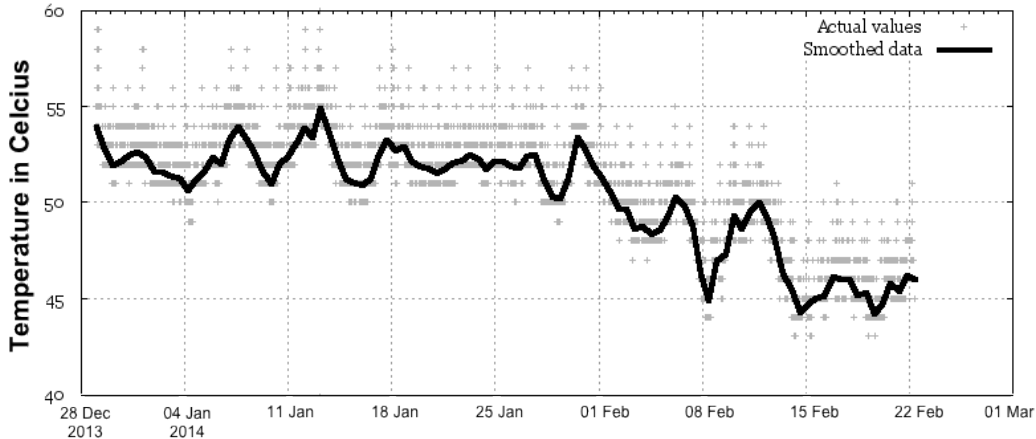


- Our eyes can deal with high data density, sinusoids, spikes, patterns, can separate noise from signal, and recognize outliers.

An example of a bad time-series plot; what problems can you identify?

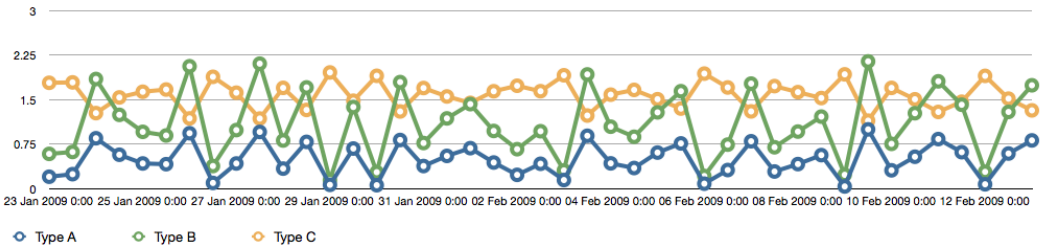


Notice how the plot's "message" is entirely different now

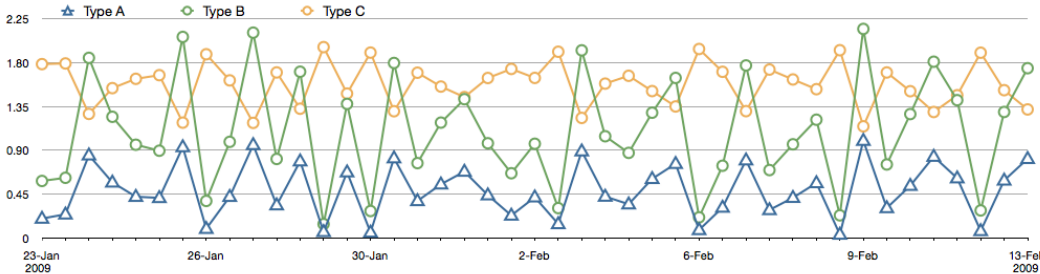


A first attempt at fixing the prior visualization

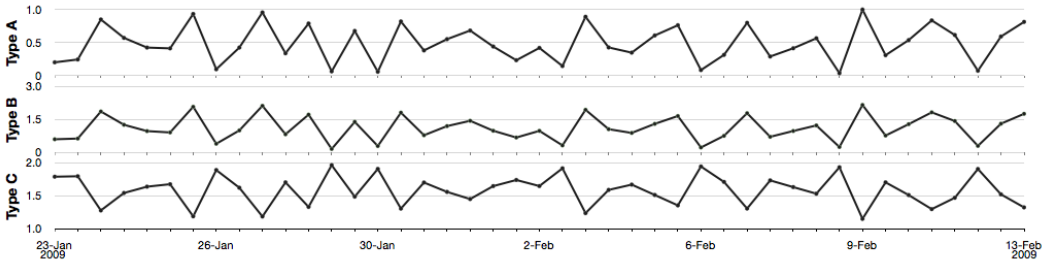
Poor plots; default settings in plotting software create cluttered plots



- The plot has been slightly improved here:



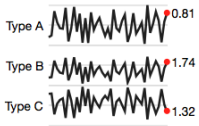
Use separate, parallel axes rather to compare plots (and minimal data ink)



These non-default settings can take a long time to set (10 minutes for this example)

Sparklines are a type of time-series plot

- except, we omit the horizontal and vertical axes (they are implicit)
- Read more about them from http://yint.org/sparklines



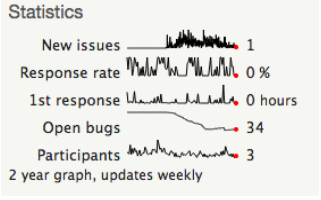
- Useful for financial trends
- Built into Excel 2010
- Great for iPods, cell phones, tablet computers
  - because they are of high density and small size
- Our eye can detect 250 dots (points) per linear inch and 650 points per square inch.

Sparklines are used on various websites now to show high-density graphics

				Valuation		
	Company name	Price	Change	Chg %	d | m | y	Mkt Cap
XFN	**iShares S&P TSX C...**	28.77	+0.04	0.14%		861.90M
HXU	Horiz. BetaPro S&...	23.23	+0.14	0.61%		65.81M
HXD	Horizns BetaPro S...	6.39	-0.04	-0.62%		68.54M
HFD	Horiz. Beta. S&P/...	3.58	+0.01	0.28%		9.39M
HEU	Hor. Beta. S&P/TS...	5.97	-0.05	-0.83%		15.95M
HFU	Horizons BetaPro ...	19.96	-0.04	-0.20%		20.40M
XEG	iShares S&P TSX C...	17.15	+0.03	0.18%		642.00M
HGU	Ho. Bta. S&P/TSX ...	7.30	-0.09	-1.22%		103.74M
HGD	Ho. Beta. S&P/TSX...	20.07	+0.26	1.31%		24.03M
XMA	iShares S&P TSX C...	12.69	+0.10	0.79%		143.53M
XTR	iShares Diversifi...	12.00	+0.03	0.25%		702.64M

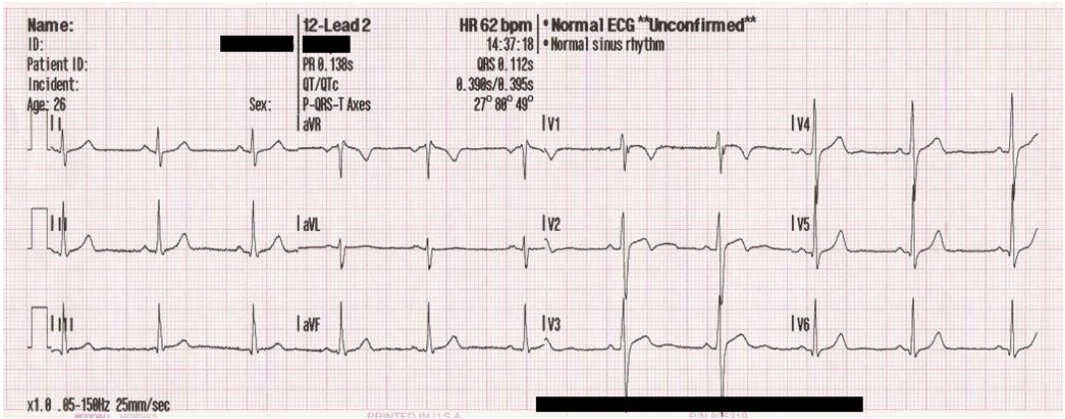
Screenshot from Google Finance 08 January 2014

Notice how you clearly detect correlations in stock prices (stocks that move together).



Screenshot from Drupal website to track software bugs.

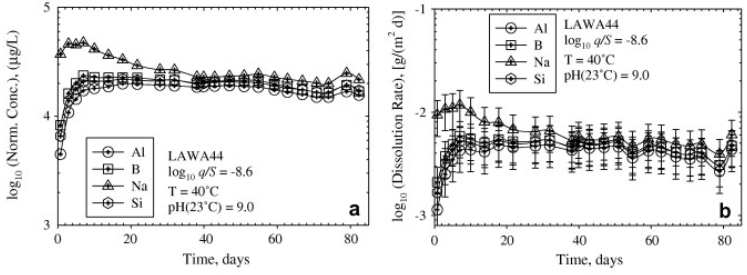
Example of sparklines in everyday use



[Wikipedia: File:12leadECG.jpg]

Keep the x-axis spacing constant on time-series plots: helps interpretation

- Use another plot (e.g. below the original) to zoom in on details



[DOI: 10.1016/j.apgeochem.2008.05.006]

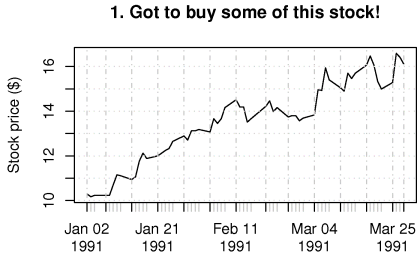
Provide an honest message to your viewer

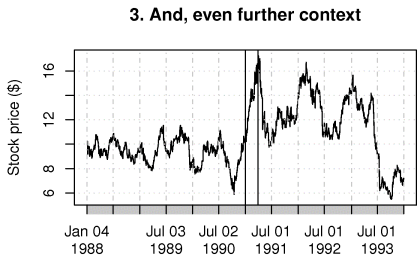
Adjust for inflation when plotting money values against time.

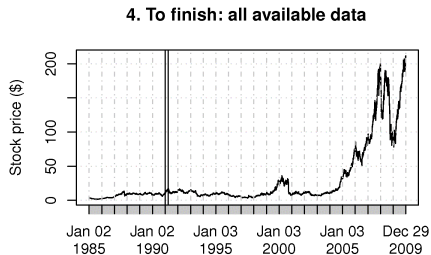
Screenshot from Google Finance.

- Example of car sales: http://www.duke.edu/∼rnau/411infla.htm

Show a reasonable amount of historical data for context



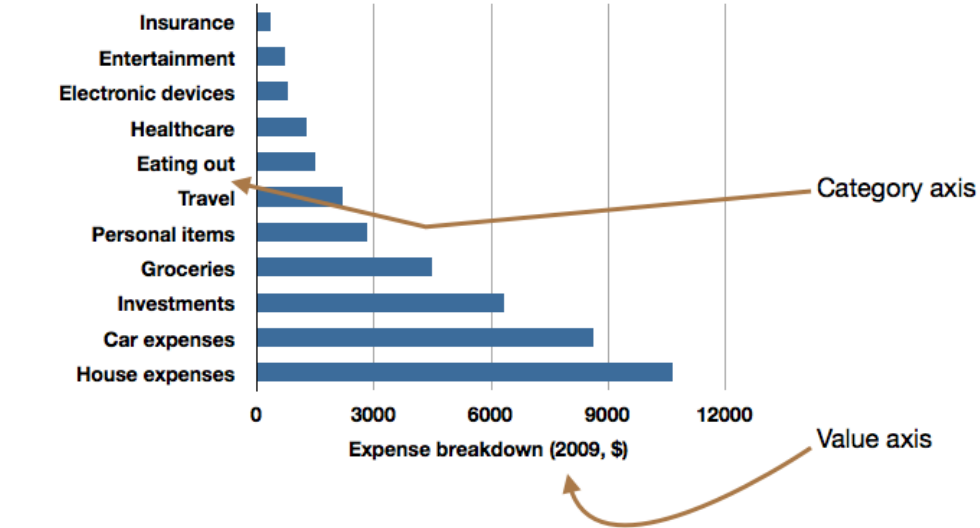




Important learning points from time-series plots

- avoid using colour as your message
- use honest scaling on your x-axis and y-axis
- the human eye (and brain) can deal with vast quantities of data: exploit it

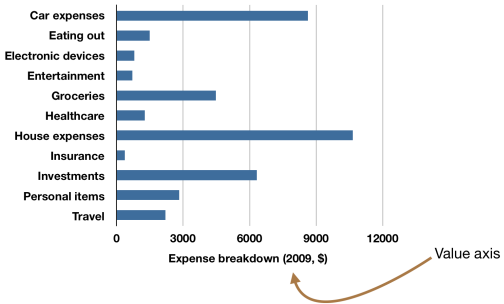
Bar plots are univariate plots, on a 2-D axis



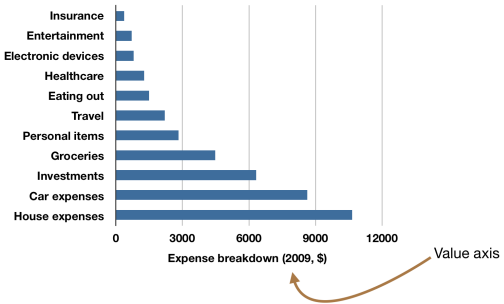
Use a bar plot when you have many categories, and the literal interpretation does not depend on category order.

Very different messages come across, even though the data are identical

There's no direct message here

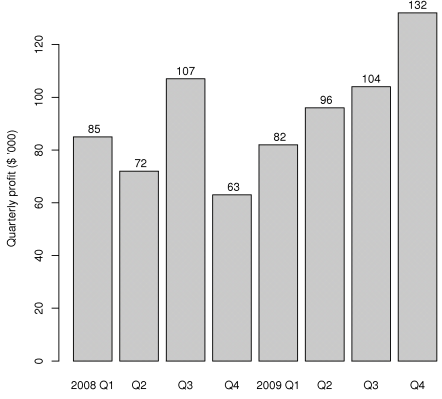


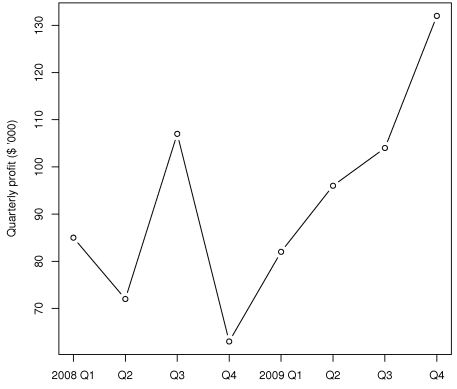
This message is more clear: the reader can quickly see their greatest expenses



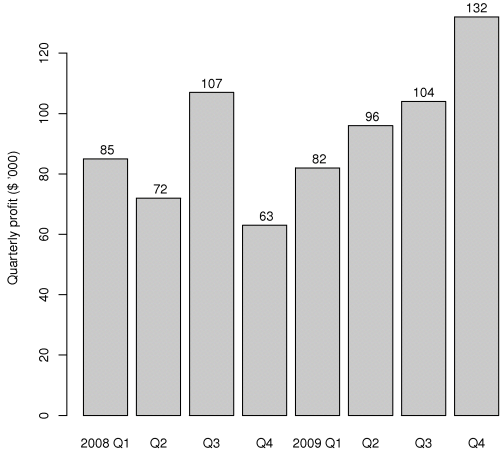
You should not use a bar plot to show time-series data

Rather use a time-series plot, which is much less wasteful and shows the trends more clearly.





Bar plots can be wasteful as each data point is repeated several times:



1. left edge (line) of each bar
2. right edge (line) of each bar
3. the height of the colour in the bar
4. the number's position (up and down along the y-axis)
5. the top edge of each bar, just below the number
6. the number itself

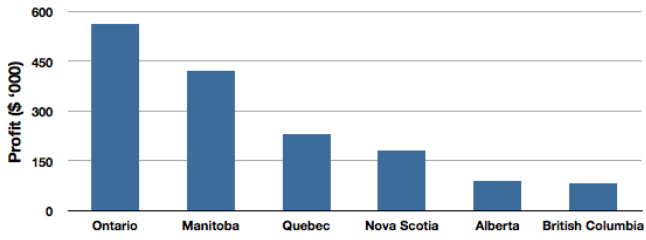
A general plotting principle: "Maximize thee data ink ratio", within reason

$$\text{Maximize data ink ratio} = \frac{\text{total ink for data}}{\text{total ink for graphics}}$$

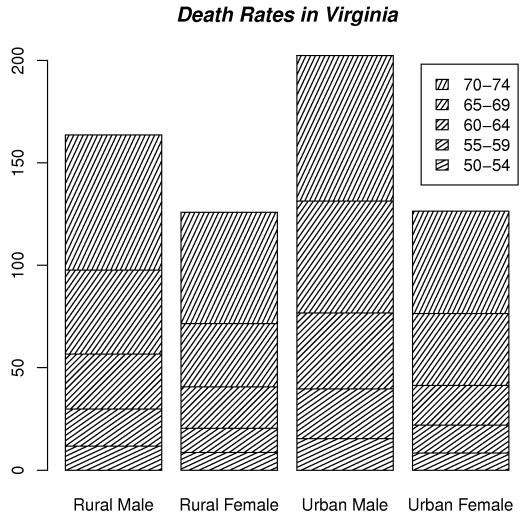
$$= 1 - \text{proportion of ink that can be erased without loss of data information}$$

For example, rather use a table for a handful of data points:

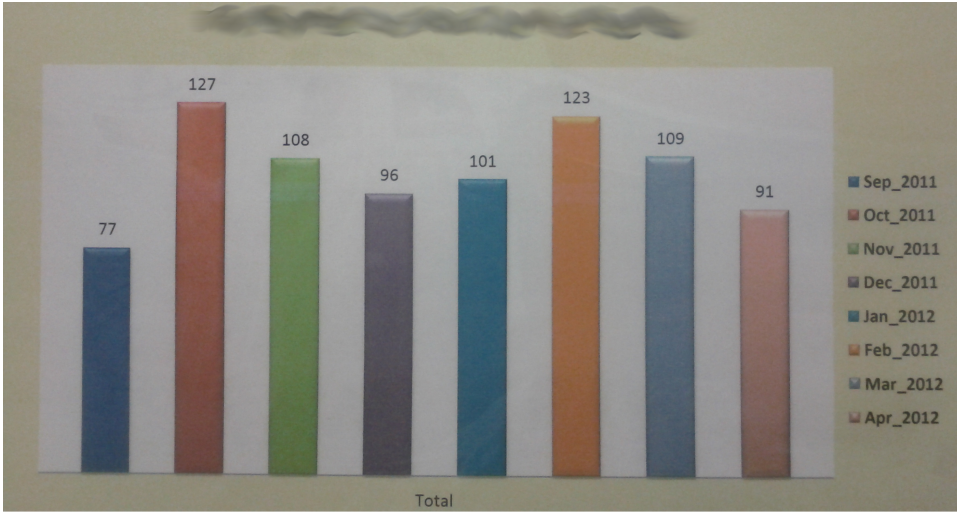
	Profit ($ '000)
Ontario	562
Manitoba	423
Quebec	231
Nova Scotia	181
Alberta	90
British Columbia	82



Don't use cross-hatching, textures, or unusual shading in the plots: it creates visual vibrations



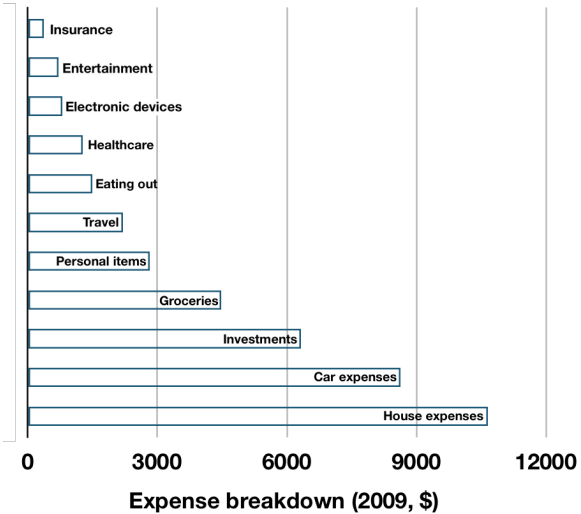
Worst bar plot ever?



Actual example from a "production report" board at a company.

Bar plots often benefit from a horizontal presentation, especially if

- there is a some ordering to the categories
- the labels do not fit side-by-side



You can place the labels inside the bars

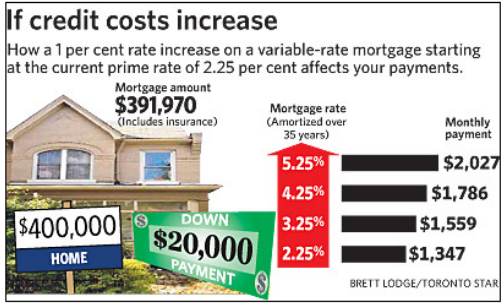
You should usually start the non-category axis at zero.

Unnecessary plot embellishments are not required

- Avoid unnecessary "extras" to enliven the plot
- "*If the statistics are boring, then you've got the wrong numbers*" [Tufte]

But living in a rented semi-detached home with three college students means she's eager to find her own space. She was also careful enough to save $40,000 for a down payment during her university years by running a College Pro Painters franchise.

Buyers today can get a variable-rate mortgage at prime or 2.25 per cent, and in many cases cheaper after discounting.



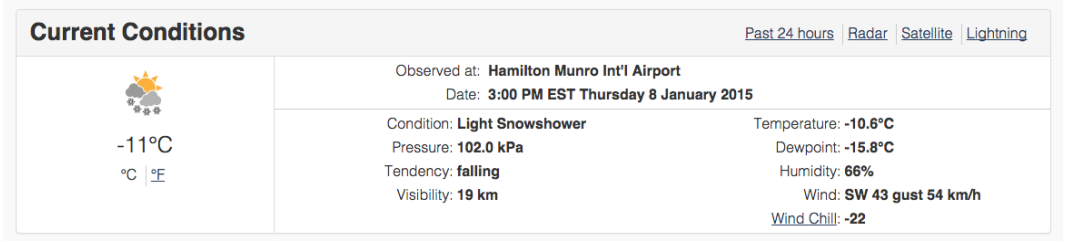
But even at the prime rate, it would cost only $1,347 to carry a $400,000 home with an amortization of 35 years and a 5 per cent down payment. By comparison, an average two-bedroom condo in the Toronto area costs $1,487 per month to rent.

That's a compelling reason for home ownership.

[Toronto Star, 2010]

Consider a vector of temperature data

For example, this is the weather in Hamilton, Ontario, on 8 January 2015



Current Conditions

Past 24 hours | Radar | Satellite | Lightning

-11°C

°C | °F

Observed at: **Hamilton Munro Int'l Airport**

Date: **3:00 PM EST Thursday 8 January 2015**

Condition: **Light Snowshower**

Pressure: **102.0 kPa**

Tendency: **falling**

Visibility: **19 km**

Temperature: **-10.6°C**

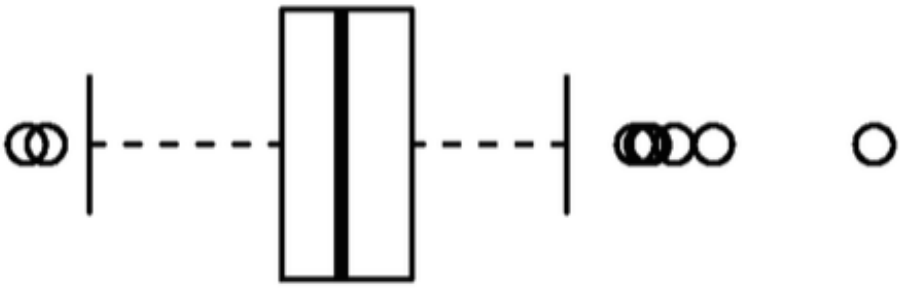
Dewpoint: **-15.8°C**

Humidity: **66%**

Wind: **SW 43 gust 54 km/h**

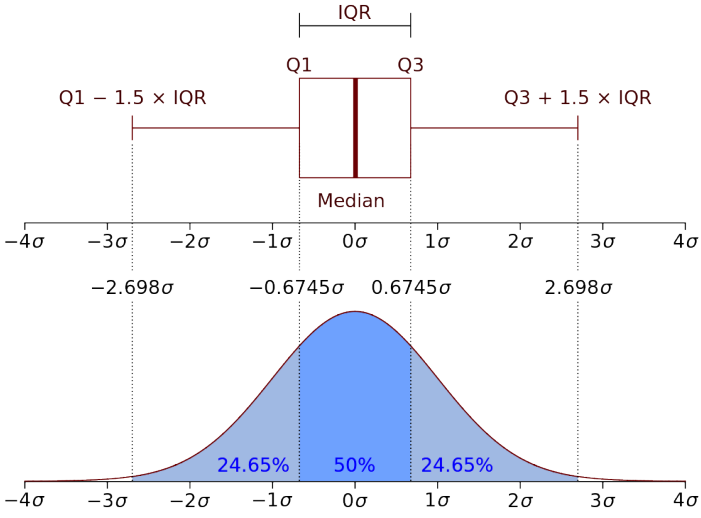
Wind Chill: **-22**

Fences and outliers illustrated on a box plot



Outliers will be defined later: they are unusual data points that are far away from the "bulk" of the data.

Box plots: compared to a pure normal distribution

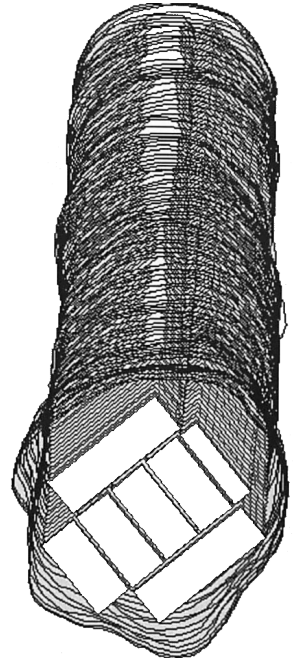


[Wikipedia has some really great illustrations to explain statistical concepts, such as this plot]

Case study: lumber cutting

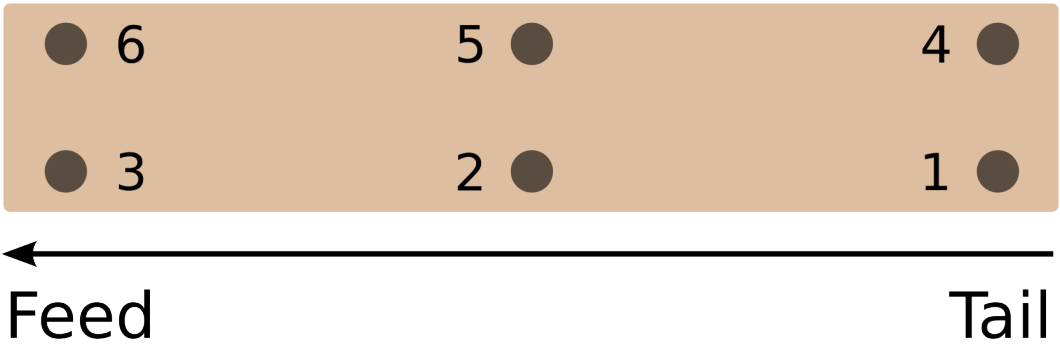
The log is completely scanned by lasers, and a few seconds later a computer determines lumber cuts that will maximize the economic value per log.

The log is rotated and guided into rigid saw blades to achieve the predicted result.



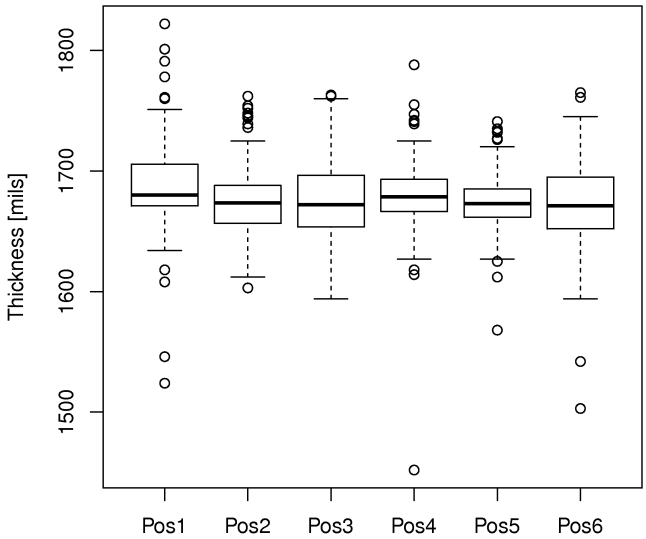
Case study: lumber cutting

After cutting, the thickness is measured at 6 locations; target = 1680 mils



Actual thickness of a 2x6 is = 1500 mils; a little extra is added to compensate for the lumber drying out

Box plots are very effective for comparing similar variables (in the same units of measurement)

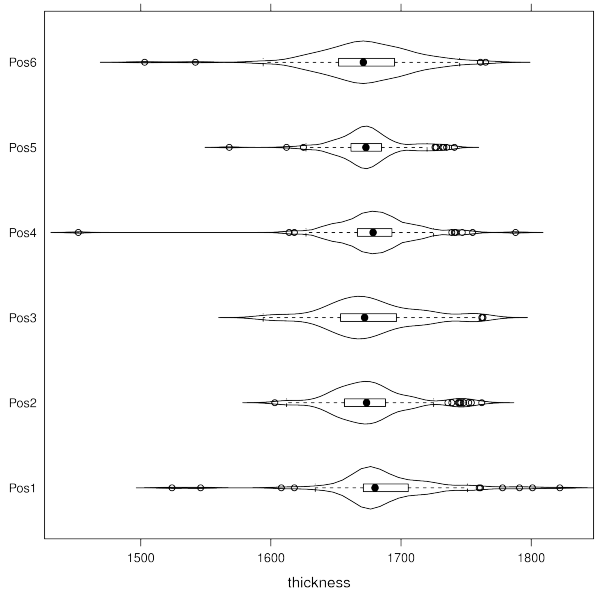


Box plots: some alternatives you might see in practice

There is no agreed on definition:

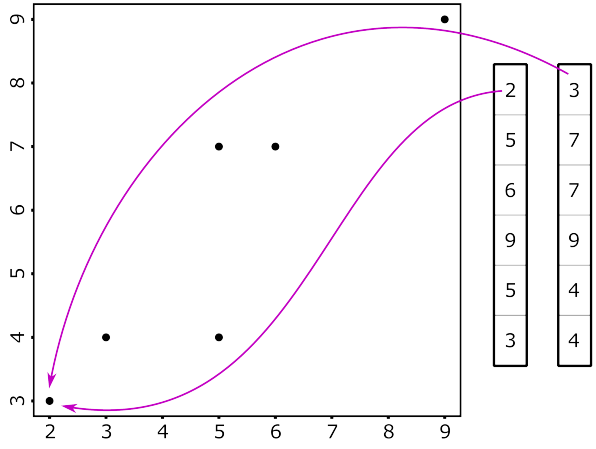
- can use the mean instead of the median
- outliers shown as dots, where an outlier is most commonly defined as any point 1.5 IQR distance units above and below the median.
- use the 2nd percentile (instead of median $- 1.5 \cdot$ IQR)
- use the 98th percentile (instead of median $+ 1.5 \cdot$ IQR)
- add the density histogram onto the box plot: *violin plot*
  - Now we can see some of the distortion at positions 1 and 3 (next slide)

Variations on a theme: the violin plot as an alternative to the box plot



Scatter plots help understand the relationship between two variables

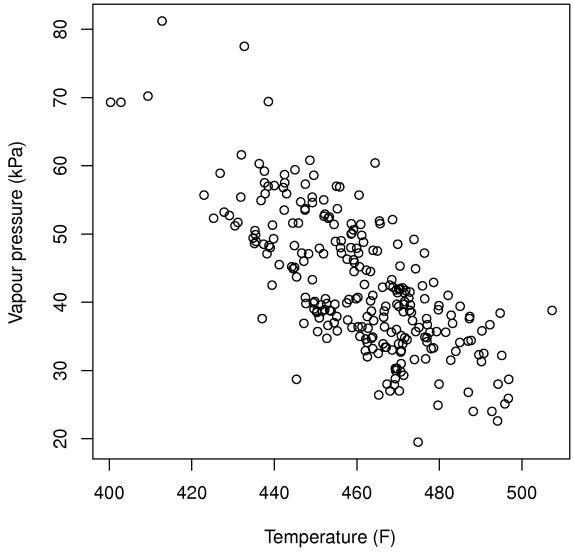
- It is a two dimensional plot of two variables (vectors).
- Each marker is the intersection of the values from the data vectors.



Intention of a scatter plot

Asks the viewer to draw a causal relationship between the two variables

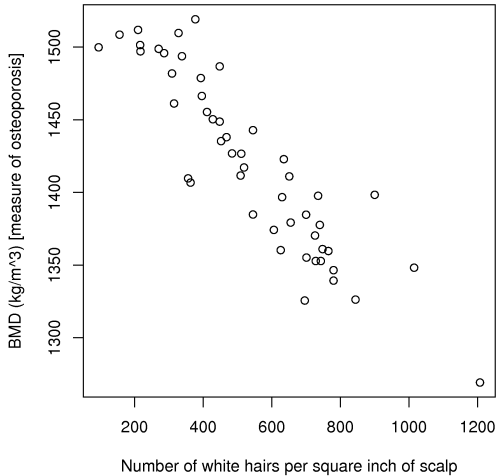
A scatter plot showing a cause-and-effect relationship



And in many cases, that causal relationship actually exists.

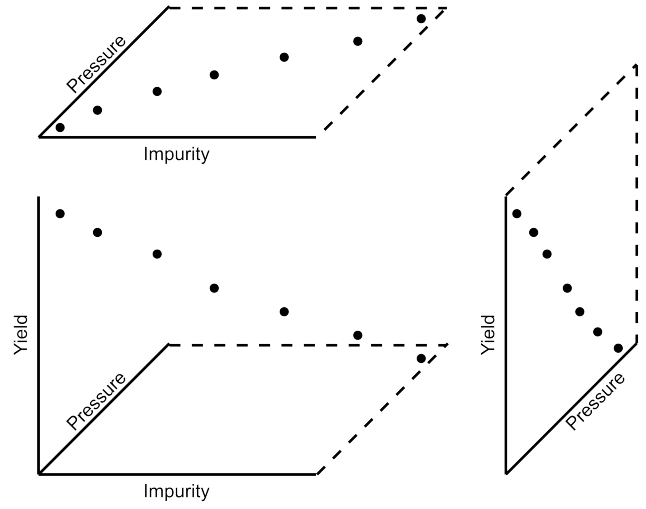
Scatter plots

However, not all scatter plots show causal phenomenon.



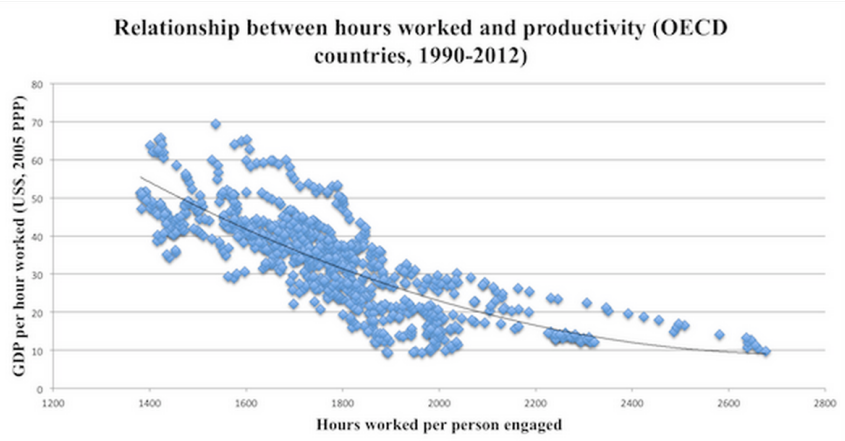
Student 2013, Hawra: "Although scatter plots may imply a cause and effect relationship exists, it is not a 'tool' to test the existence of a possible relationship."

Three variables: so 3 scatter plot combinations could have been drawn. Which are correlations, and which are actually cause-and-effect?



We will answer the question in the Experiments section.

Scatter plots: is there cause and effect here?

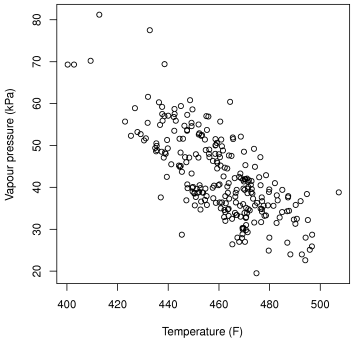


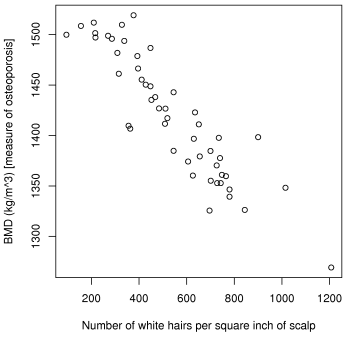
http://yint.org/working-hours explains the relationship and the data source.

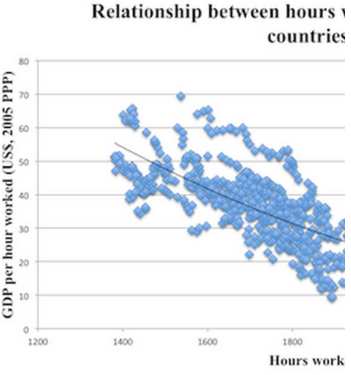
Scatter plots can be greatly improved from the software defaults by:

- making each axis as tight as possible
- avoiding heavy grid lines
- use the least amount of ink
- not distorting the axes

Note the tight axes, low amount of data ink: scatter plots are efficient.







There is an unfounded fear that others won't understand your scatter plot

- Plant control room: seldom see scatter plots.
- Tufte study (VDQI, http://yint.org/vdqi): no scatter plots in a sample of Western daily newspapers (1974 to 1980)
- Japanese newspapers frequently use scatterplots
- He shows 12 year olds can interpret such plots.

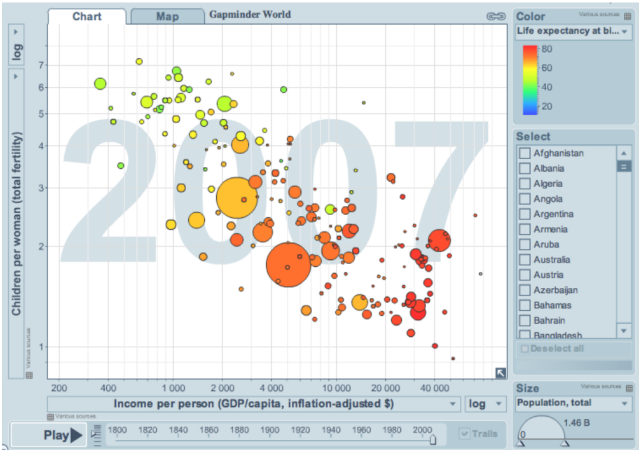
Key point

The producers of charts must assume their audience is capable of interpreting them.

Rather, assume that if you can understand the plot, so will your audience.

Take a diversion to watch this YouTube video http://yint.org/rosling-video

Hans Rosling has used these to great effect to illustrate issues related to International Health



Variables shown in the figure:

1. x-axis: income per person
2. y-axis: children per woman
3. marker area: population
4. colour: life-expectancy [20 to 80]
5. time-based animation: on the GapMinder website you can "play" the graph over time

The http://gapminder.org site allows you to select many interesting variables on the axes.

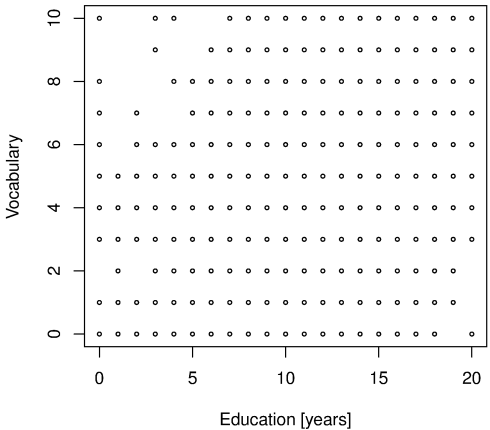
Watch a video explanation of these data: http://yint.org/rosling-video

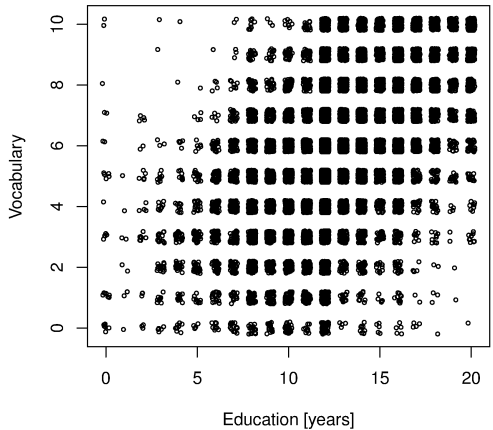
3D glasses used to visualize process data in 6-dimensions?

It would be possible with 3D glasses.

[Flickr: #8443551051]

Scatter plots lose density information: recover it with some jitter



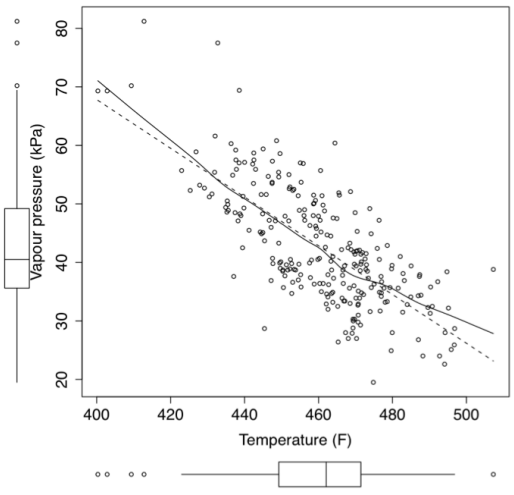


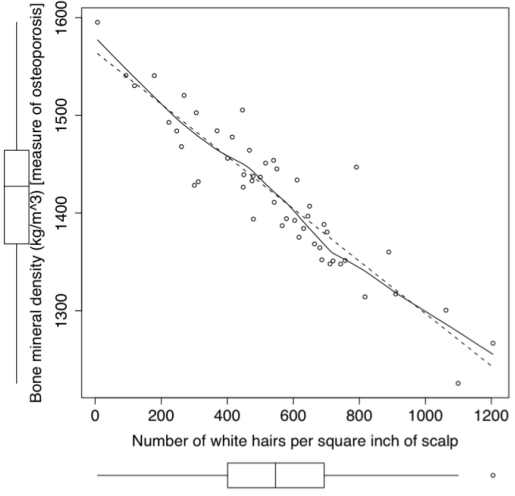
R code:

```
plot(education, vocabulary)
```

```
plot(jitter(education), jitter(vocabulary))
```

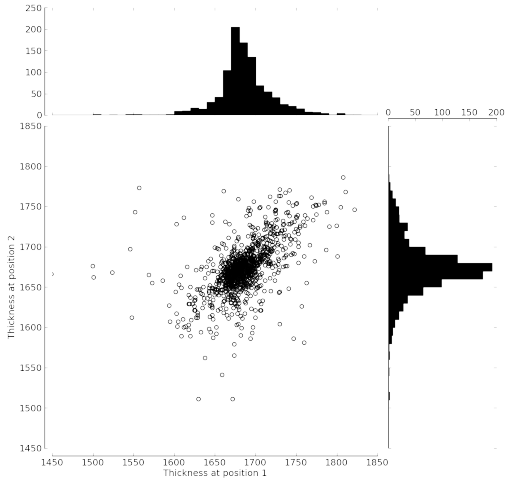
Recover distribution and spread information with box plots



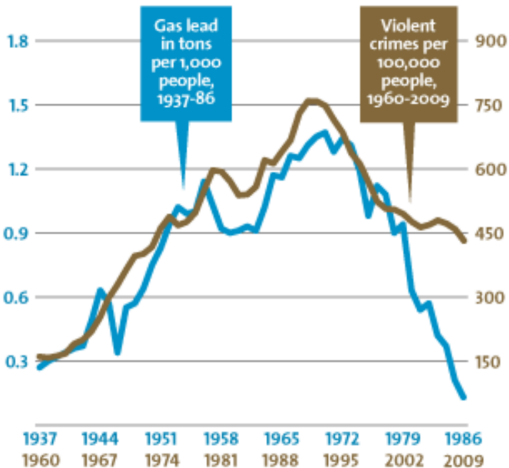


Used the `scatterplot(...)` function from the `library(car)` in R to create these.

Consider adding histograms when you are exploring the data to learn about the system

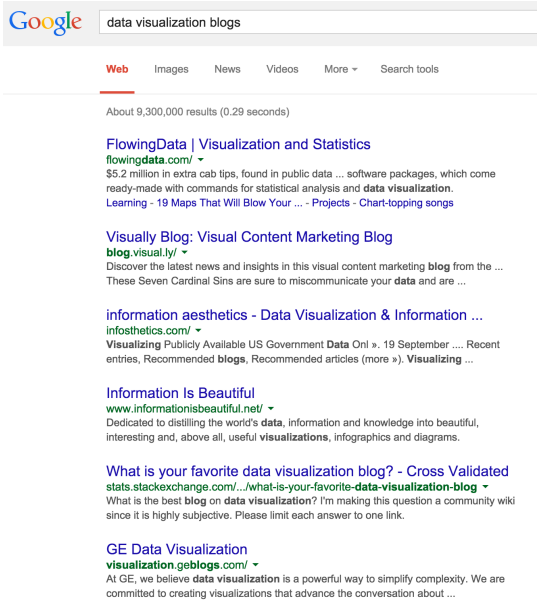


Investigate this plot in your own time



- Why did the author use a time-series plot to show correlation?
- Would the plot be more informative as 2D-scatter plot?
- Redraw a rough version of this plot as a scatter plot instead.
- What if you were to repeat this analysis for multiple regions/countries/cities. How would you show (visualize) the correlations effectively?
- Read the full article for details: http://yint.org/lead-and-crime

Some data visualization blogs worth reading



Tables

Tables are for **comparative** data analysis on **categorical objects**.

	Bank loan monthly payments	Monthly lease payment	Minimum downpayment for lease	Total interest paid over 48 months	Monthly insurance payment
Ford Fusion	552	395	0	2,529	180
Honda Civic	538	424	0	2,466	236
Mazda 3	506	478	1,000	2,318	251
Toyota Yaris	435	490	1,000	1,992	198
VW Golf	596	550	2,500	2,730	244

- **categorical objects**: the cars
- Note the rows are in *default* alphabetical order.
- We can make the table "tell a story" if we reorder the rows by some other variable.
  - e.g. monthly insurance payment

Tables

- Compare defect types (columns) for different product grades (rows)
- Categorical variables appear in the **rows** and **columns** here

	Total defects	A	B	C	D	E
A4636	131	37	21	28		45
A2524	86	20	24	21	1	20
A3713	75	17	13	18		27
A4452	73	5	33	17		18
A4088	72	14	16	12	2	28
A2103	68	14	13	14	1	26
A2156	68	16	13	19	2	18
A3681	66	12	16	9	1	28
A1366	50	11	15	12		12
A2610	39	5	7	12		15
Total	**728**	**151**	**171**	**162**	**7**	**237**

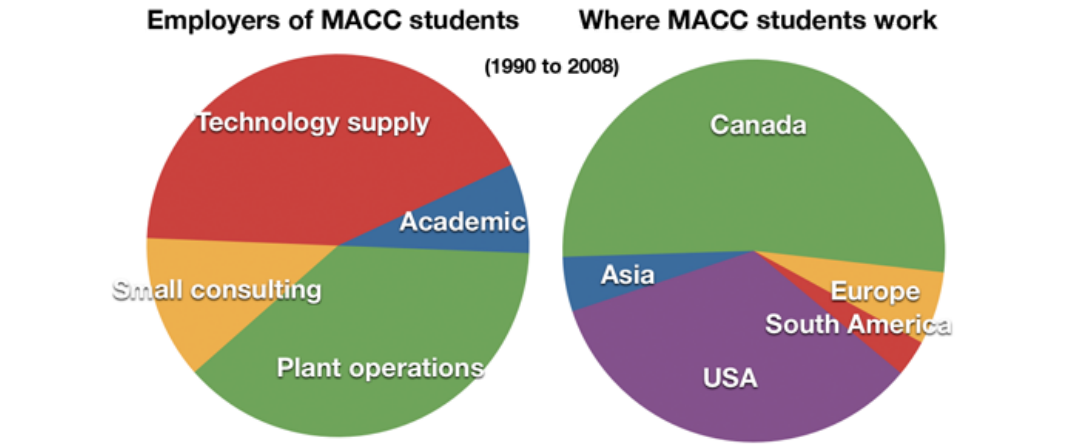
- Which defects cost us the most money?

Tables

- Defect frequency
  - If 1850 lots of grade A4636 (first row): defect A rate $= 1/50$
  - If 250 lots of grade A2610 (last row): defect A rate $= 1/50$
  - Redraw table on production rate basis
- If comparing defects over different grades: go down the table (show fraction within the column)
- If comparing defects within grade: go across table (show fraction with the row)
  - Could weight each column by cost of defect

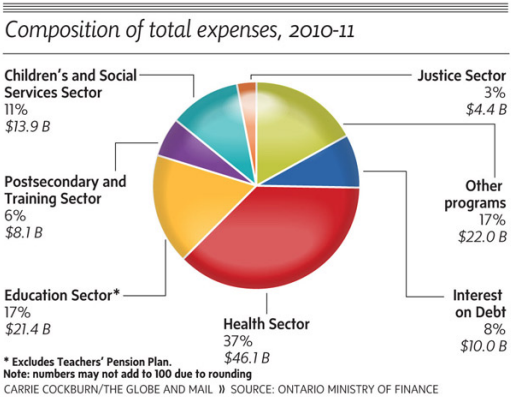
Tables

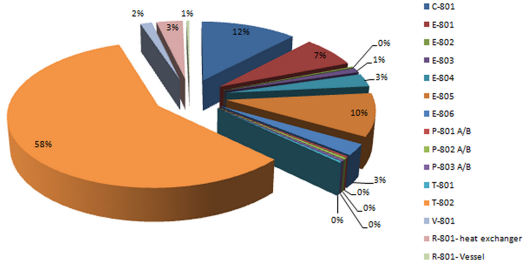
Common pitfalls: using pie charts when tables will do

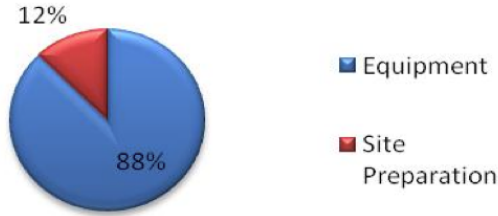


I cannot explain the pitfalls of pie charts as well as Stephen Few does: Save the pies for dessert (please read)

Tables vs pie charts: plenty of bad examples







Tables

2. arbitrarily ordering of the rows

	Bank loan monthly payments	Monthly lease payment	Minimum downpayment for lease	Total interest paid over 48 months	Monthly insurance payment
Ford Fusion	552	395	0	2,529	180
Honda Civic	538	424	0	2,466	236
Mazda 3	506	478	1,000	2,318	251
Toyota Yaris	435	490	1,000	1,992	198
VW Golf	596	550	2,500	2,730	244

Tables

3. using excessive grid lines

	Total defects	A	B	C	D	E
A4636	131	37	21	28		45
A2524	86	20	24	21	1	20
A3713	75	17	13	18		27
A4452	73	5	33	17		18
A4088	72	14	16	12	2	28
A2103	68	14	13	14	1	26
A2156	68	16	13	19	2	18
A3681	66	12	16	9	1	28
A1366	50	11	15	12		12
A2610	39	5	7	12		15
Total	**728**	**151**	**171**	**162**	**7**	**237**

	Total defects	A	B	C	D	E
A4636	131	37	21	28		45
A2524	86	20	24	21	1	20
A3713	75	17	13	18		27
A4452	73	5	33	17		18
A4088	72	14	16	12	2	28
A2103	68	14	13	14	1	26
A2156	68	16	13	19	2	18
A3681	66	12	16	9	1	28
A1366	50	11	15	12		12
A2610	39	5	7	12		15
Total	**728**	**151**	**171**	**162**	**7**	**237**

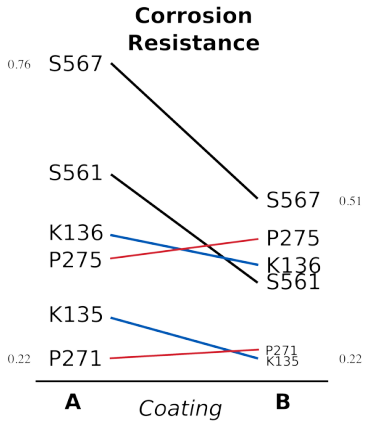
Tables

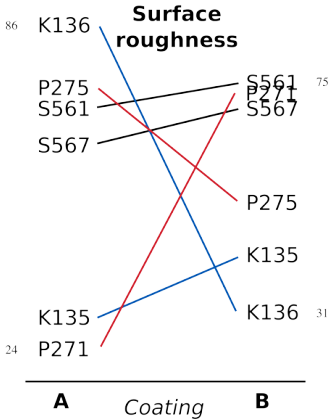
Interesting example: comparing two treatments

	Corrosion resistance		Surface roughness	
	A	B	A	B
K135	0.3	0.22	30	42
K136	0.45	0.39	86	31
P271	0.22	0.24	24	73
P275	0.4	0.44	74	52
S561	0.56	0.36	70	75
S567	0.76	0.51	63	70

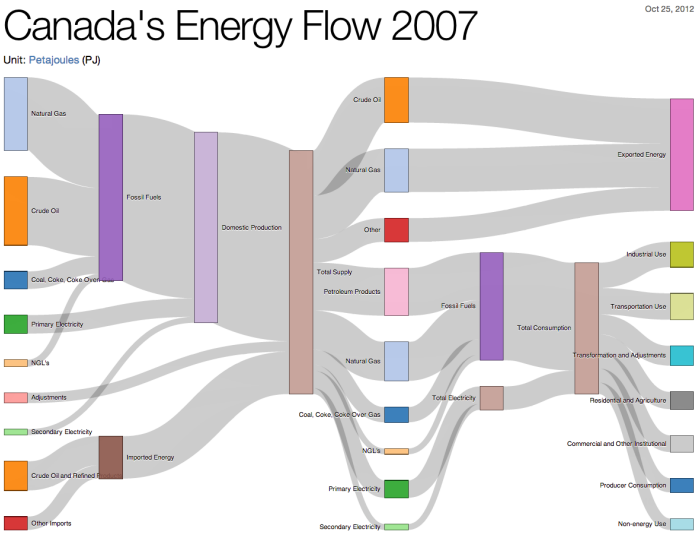
- Coating A or B are applied to different products
- K-series, P-series, S-series
- How does the coating affect corrosion and surface roughness?

Tables


Corrosion Resistance
0.76
S567
S561
K136
P275
K135
0.22
P271
S567
0.51
P275
K136
S561
P271
K135
0.22
A
Coating
B


Surface roughness
86
K136
P275
S561
S567
24
P271
S561
P271
S567
75
P275
K135
K135
K136
31
A
Coating
B

Sankey diagrams



[http://gnewton.ca/gn/sankey/sankey_canada_energy_flow_2007.html]

Aesthetics and style

I highly recommend reading Tufte's 4 books: contain remarkable examples of how to bring data to life.

Colour

- Colour is effective, but:
  - readers could be colour-blind,
  - document read from a gray-scale print out
- There is **no standard colour progression** (blues, greens, yellows, orange, red).
- Safest colour progression is gray-scale axis: from black to white
  - satisfies colour-blind readers
  - looks good in printed form

General summary

No general advice that applies in every instance. Useful tips nevertheless:

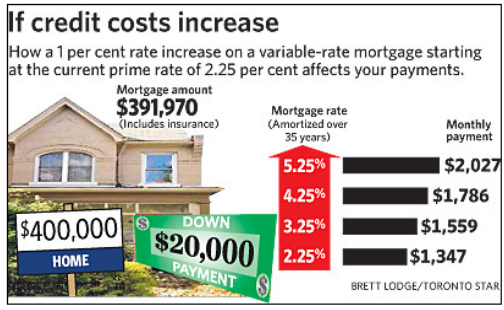
- To understand causality, you must show causality: use bivariate scatter plots (sometimes line plots also work well)
- Plots and text go together: a plot = paragraph of text
  - add labels to plots for outliers and interesting points
  - add equations
  - add small summary tables
- Avoid codes: "A = grade TK133", "B = grade RT231"

General summary

- Avoid unnecessary "extras" to enliven the plot
- "*If the statistics are boring, then you've got the wrong numbers*".

But living in a rented semi-detached home with three college students means she's eager to find her own space. She was also careful enough to save $40,000 for a down payment during her university years by running a College Pro Painters franchise.

Buyers today can get a variable-rate mortgage at prime or 2.25 per cent, and in many cases cheaper after discounting.



But even at the prime rate, it would cost only $1,347 to carry a $400,000 home with an amortization of 35 years and a 5 per cent down payment. By comparison, an average two-bedroom condo in the Toronto area costs $1,487 per month to rent.

That's a compelling reason for home ownership.

General summary

- Adjust for inflation if plot involves money and time
- Maximize the data-ink ratio = (ink for data) / (total ink for graphics).
  1. eliminate non-data ink
  2. erase redundant data-ink.
- Maximize data density: 250 data points per linear inch, and 625 data points per square inch.

General summary

Good plotting is not difficult. It just takes time and thought.