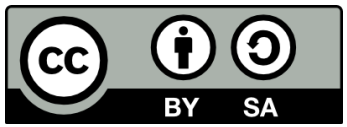
Statistics for Engineers

ChE 4C3 and 6C3

kevin.dunn@mcmaster.ca
http://learnche.mcmaster.ca/4C3

Overall revision number: 19 (January 2013)

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

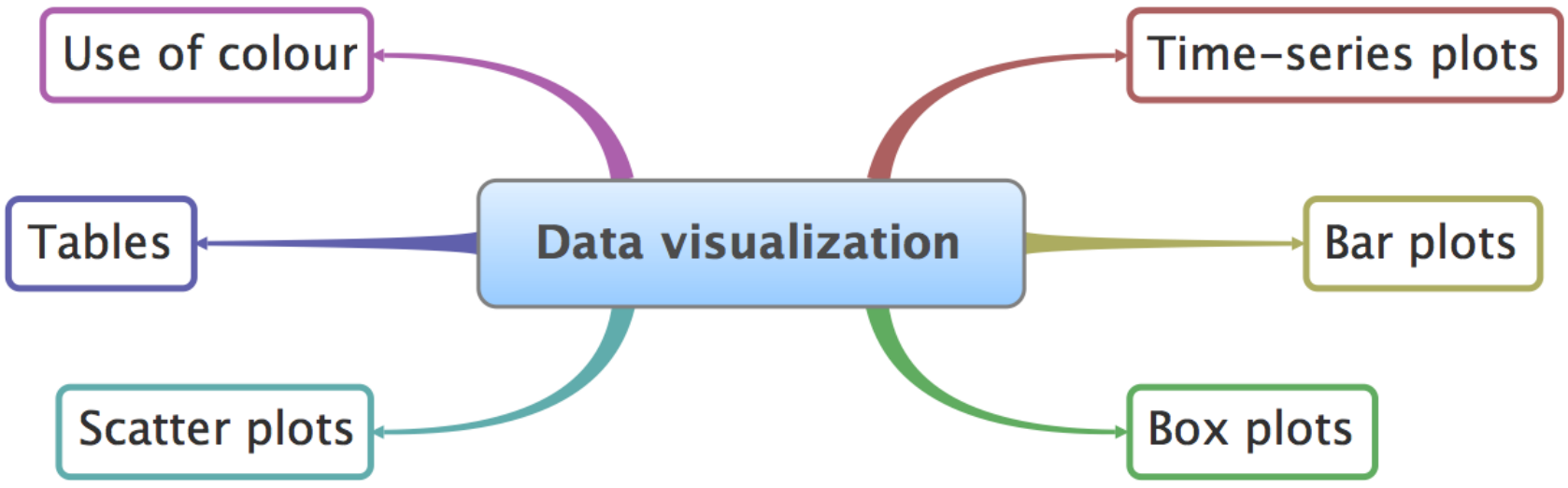
`kevin.dunn@mcmaster.ca`

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: 19

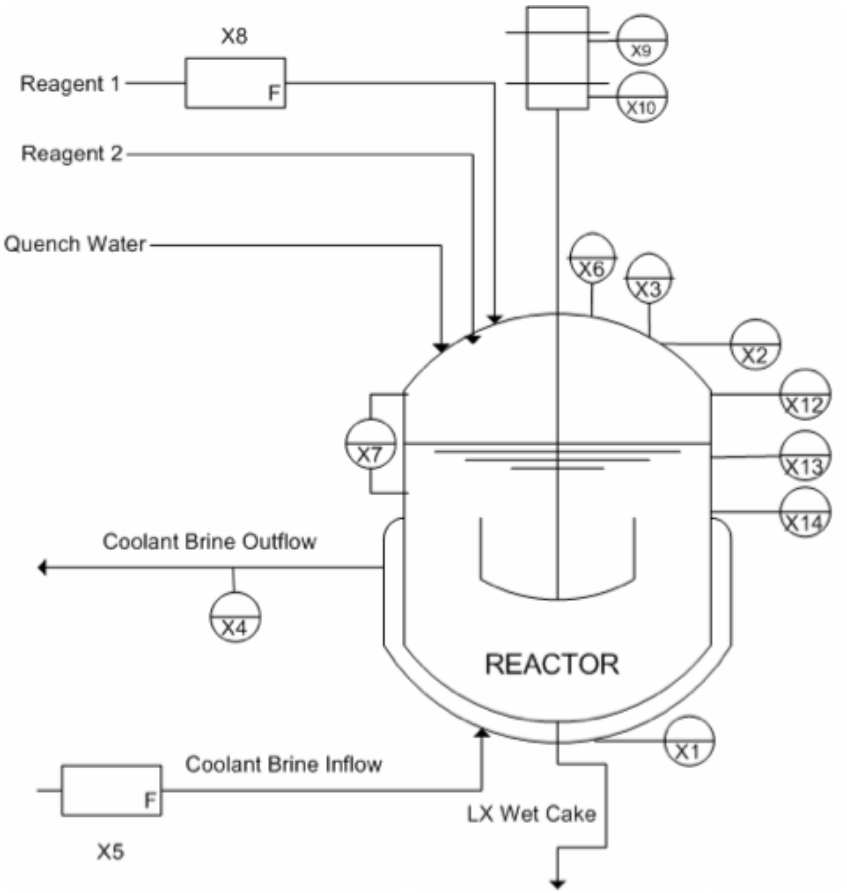
Plot your data



Usage examples

- *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
- Tiffany's example
- *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: large quantities of valuable data

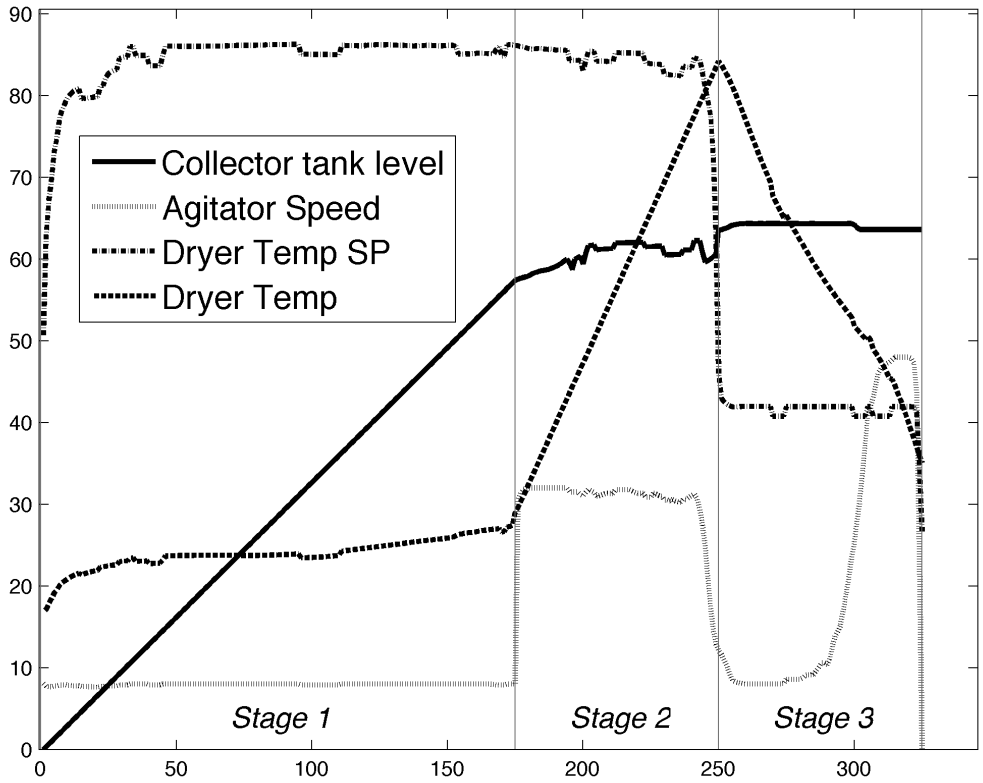


[From Cecilia Rodrigues' M.A.Sc thesis, 2006, McMaster University, used with permission]

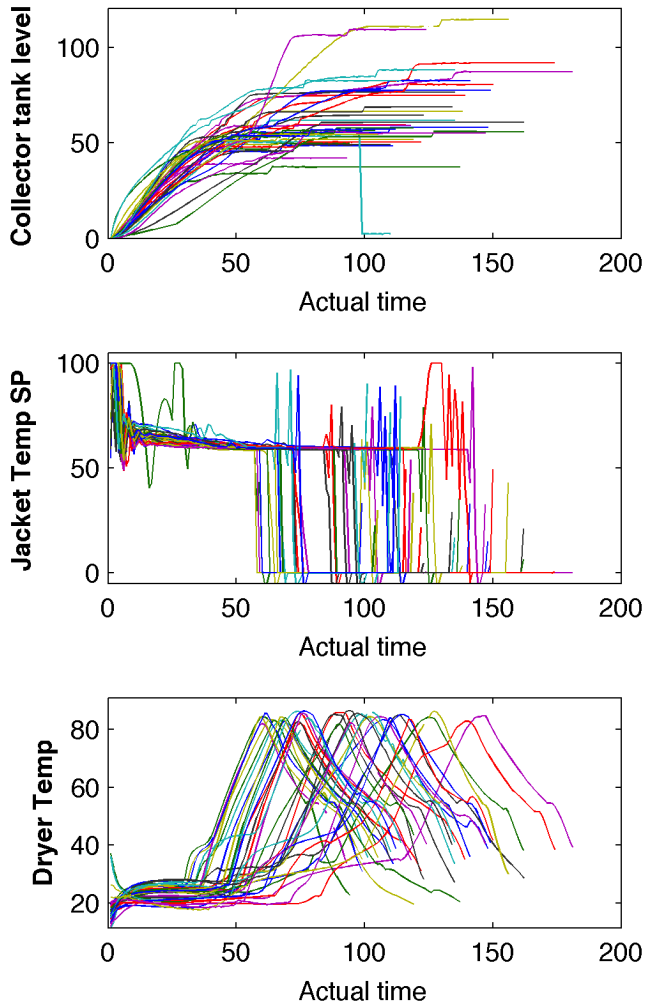
[Flickr: #2516220152]

Batch systems: large quantities of valuable data

Data from a single batch



Data from many batches



References

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

Background

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"
- We will look at: **how** and show examples of bad plots

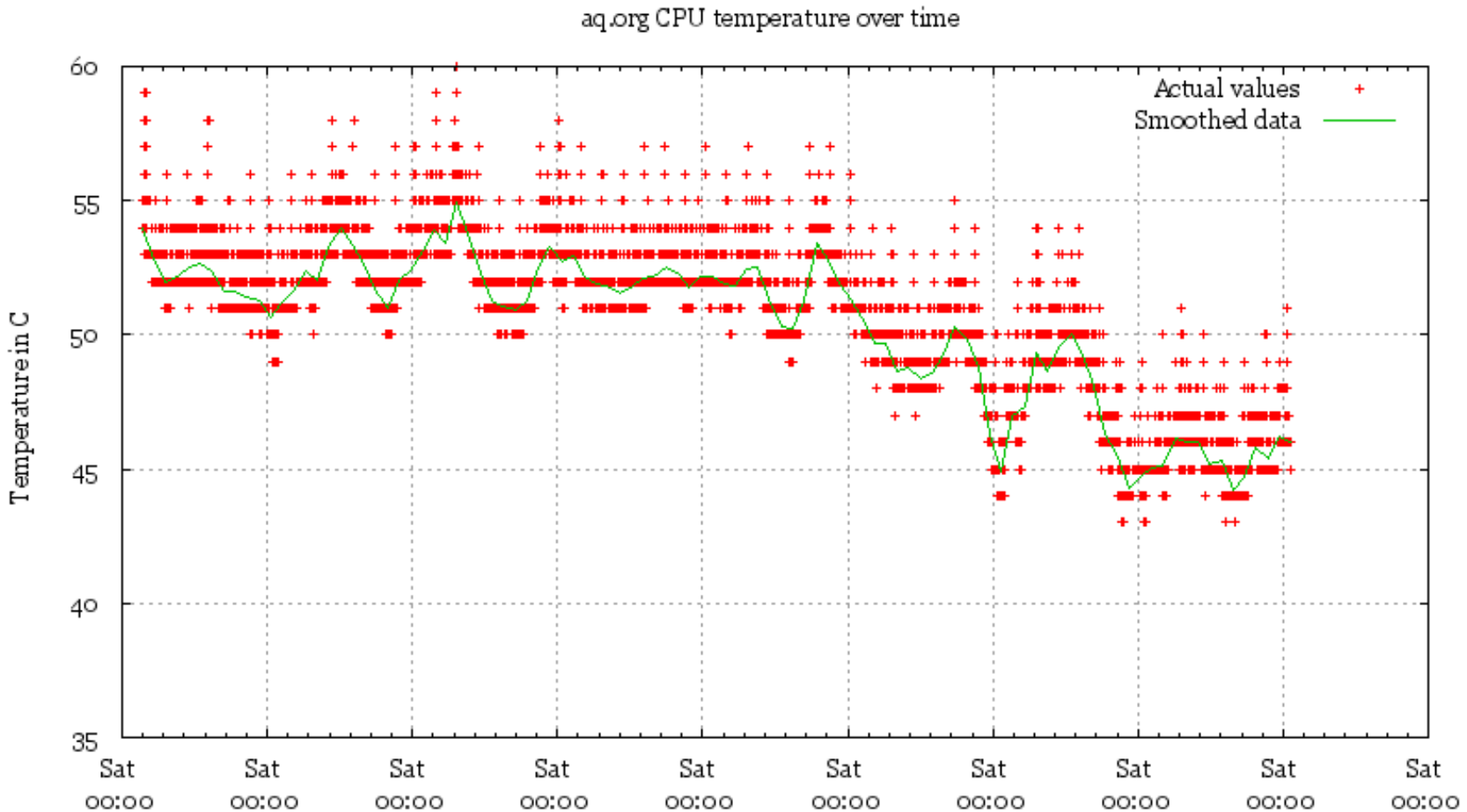
Time-series plots

- It is a 2-dimensional plot:
  - (usually) horizontal x-axis: time or sequence order
  - other axis: the data values
- Univariate plot
- Our eyes can deal with high data density:
  - sinusoids
  - spikes
  - outliers
  - separate noise from signal

Time-series plots

Good, automated labelling is important.

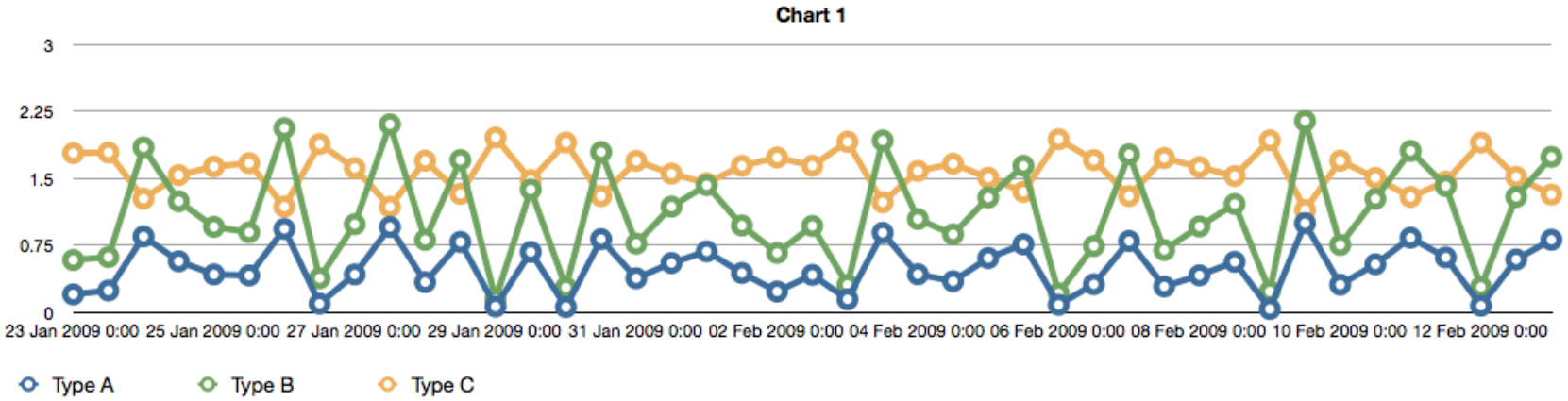
Here's an example of bad labelling



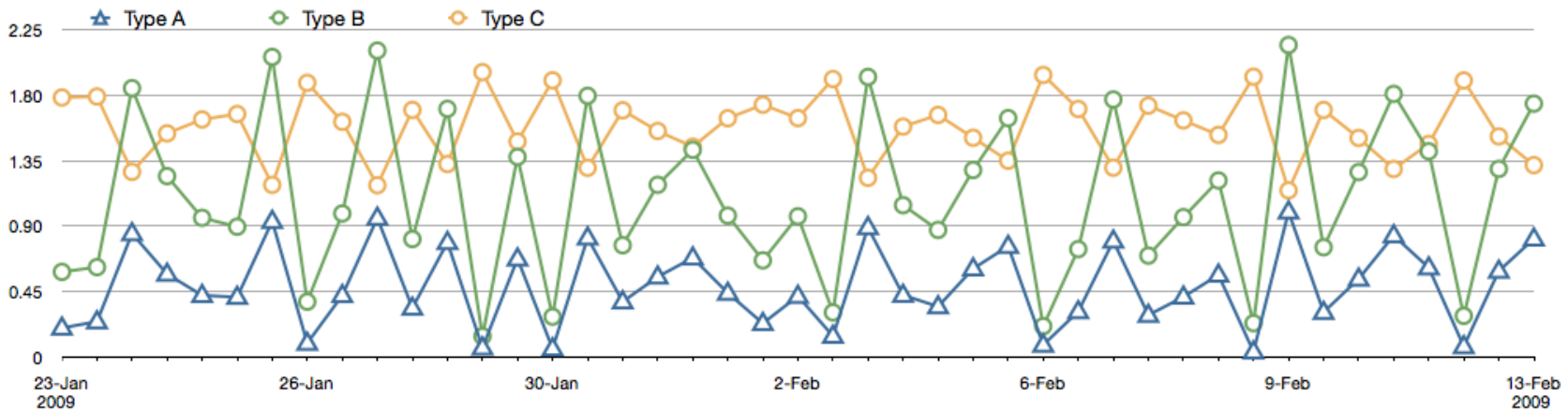
(and bad axis scaling and colour choices)

Time-series plots

- Multiple lines (trajectories): should not cross and jumble

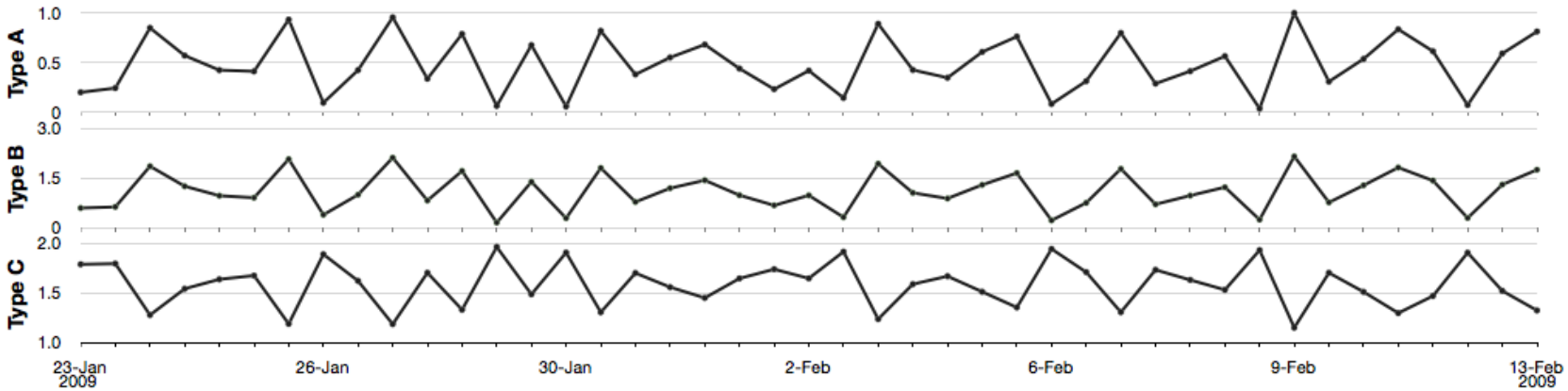


- Colours and markers help only slightly



Time-series plots

Use separate, parallel axes rather; and minimal ink

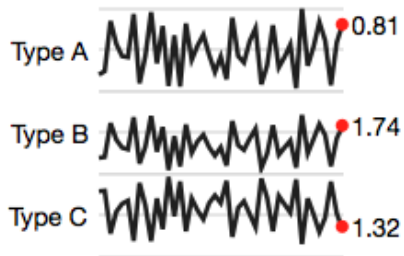


These non-default settings can take a long time to set (10 minutes for this example)

Time-series plots

Sparklines

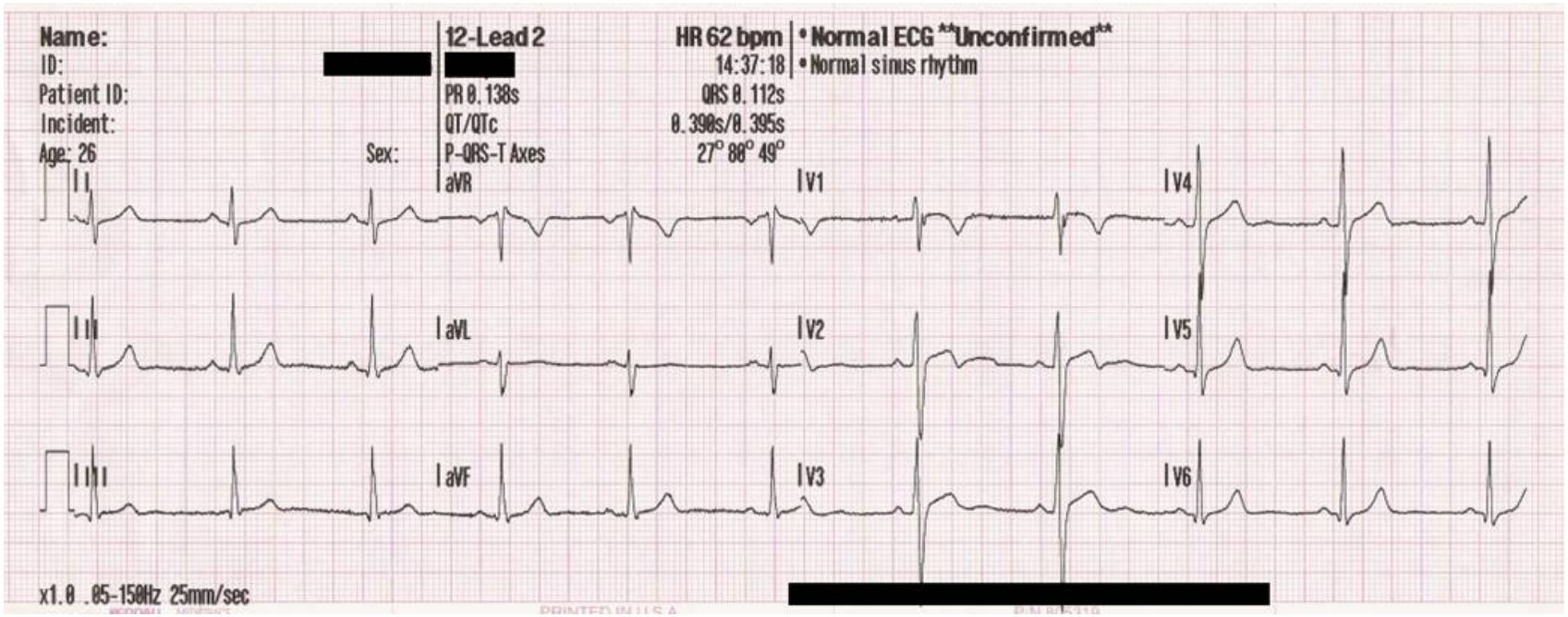
- Read more about them from this website (link also in the notes)



- Used for financial trends (see Google Finance, for example)
- Built into Excel 2010
- Good for iPods, cell phones, tablet computers:
  - high density, small size.

Time-series plots

Example of sparklines in everyday use:

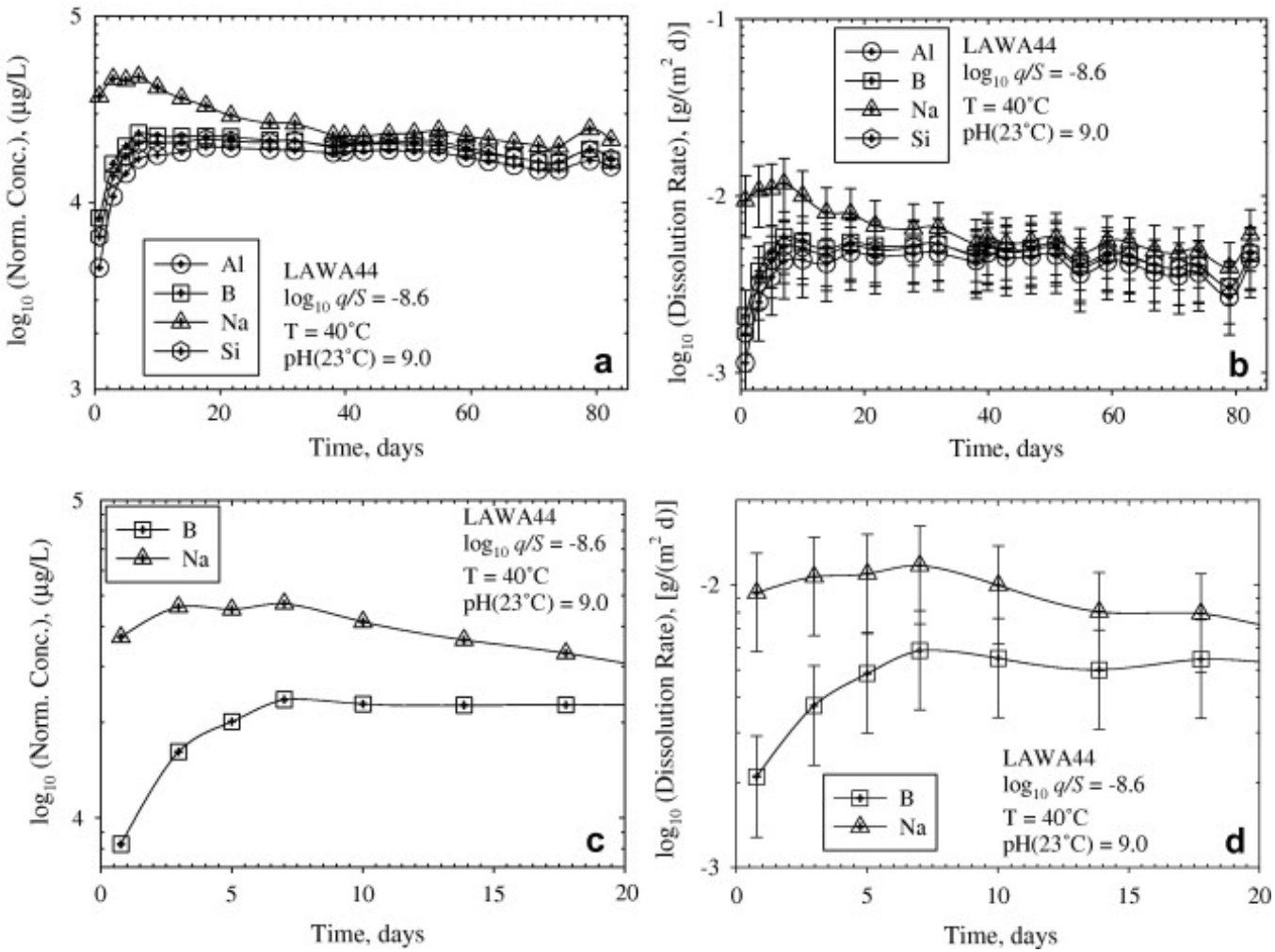


[Wikipedia: File:12leadECG.jpg]

Time-series plots

Further tips

- Keep the x-axis spacing constant: helps interpretation
  - Keep constant spacing on a time-axis (months)
  - Don't use magnifying glass concept; rather show a second plot



[DOI: 10.1016/j.apgeochem.2008.05.006]

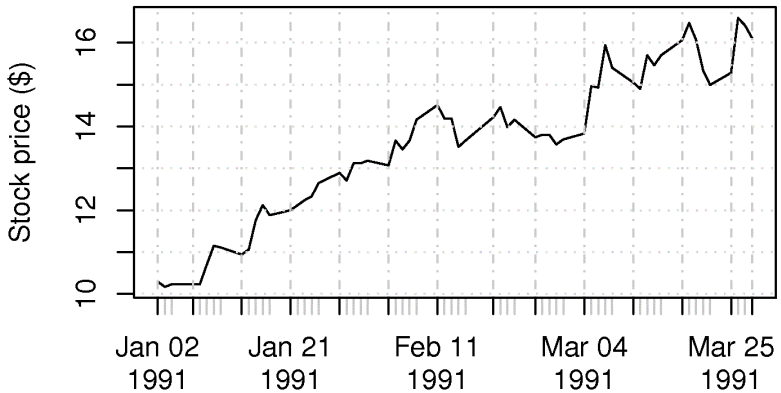
Time-series plots

- Adjust for inflation when plotting money values against time
  - sales of polymer to DuPont over the past 10 years
  - example of car sales:
    http://www.duke.edu/∼rnau/411infla.htm

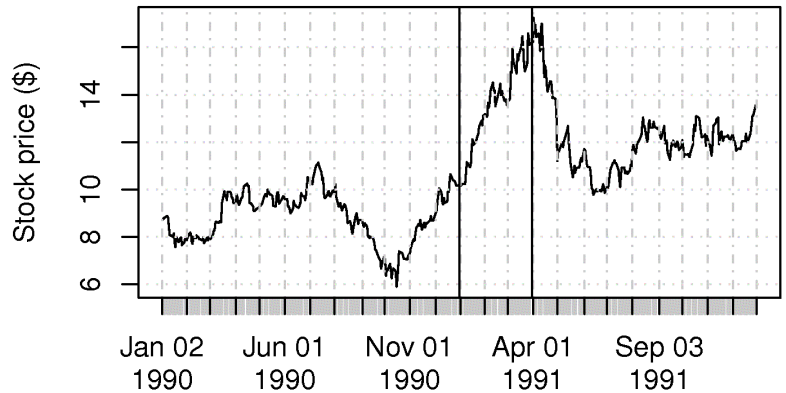
Time-series plots

Show reasonable amount of data for context

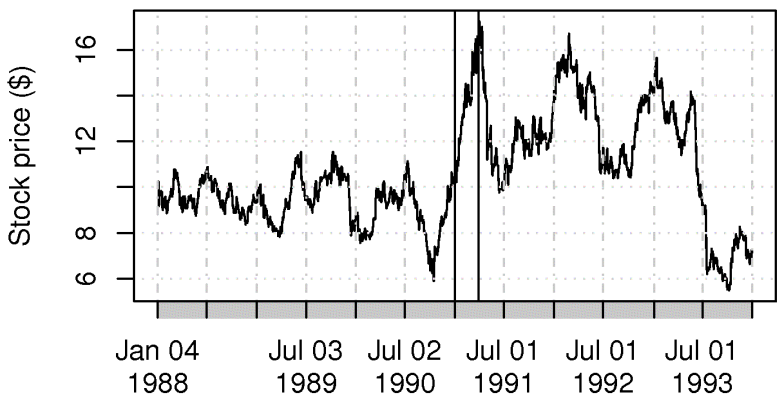
1. Got to buy some of this stock!



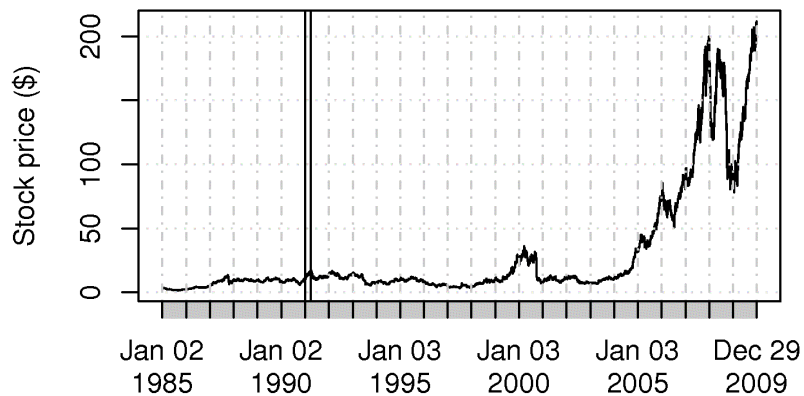
2. But, here is some more context



3. And, even further context

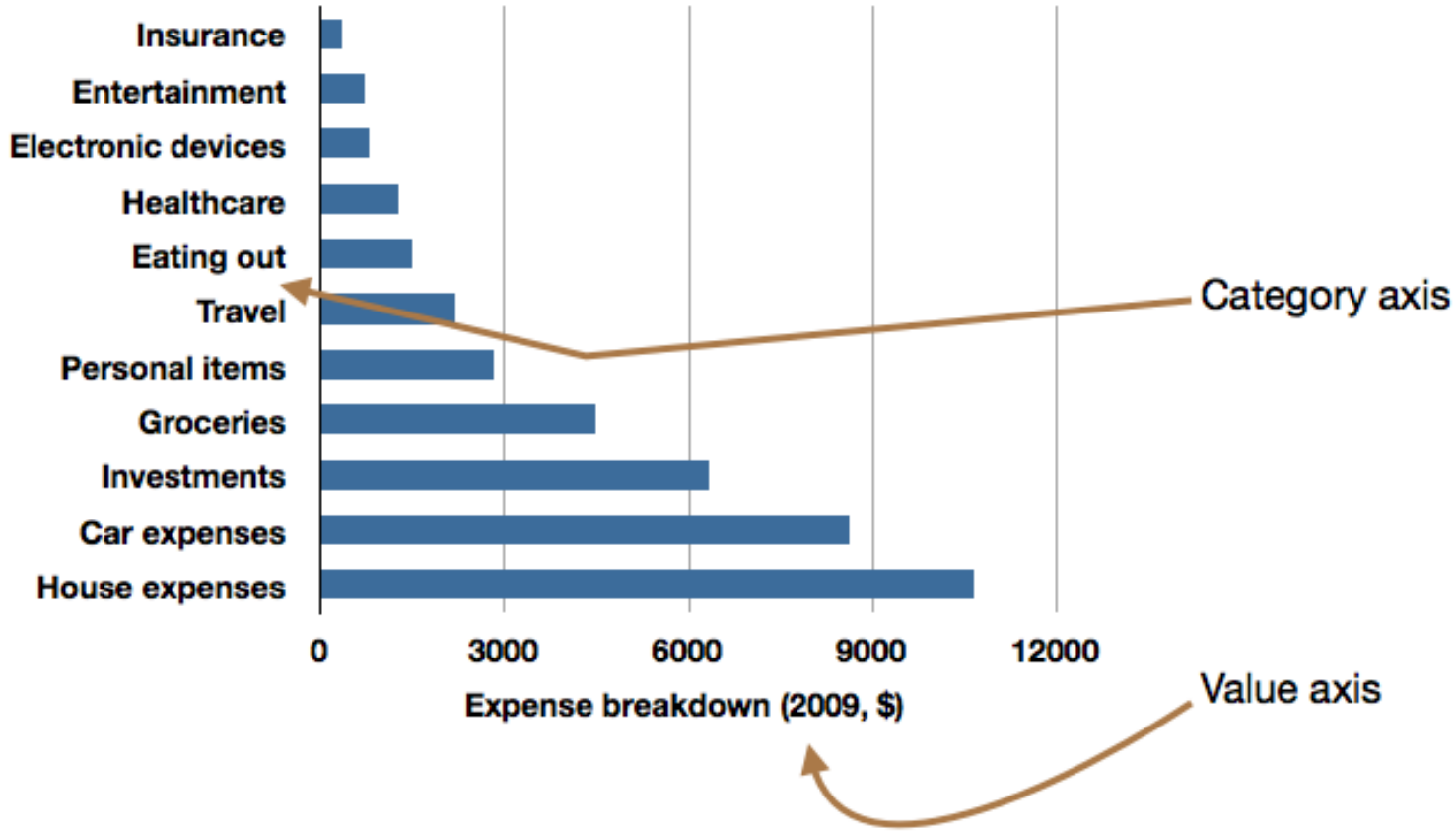


4. To finish: all available data



Bar plots

- A univariate plot on a two dimensional axis.
- Has a *category axis* and *value axis*

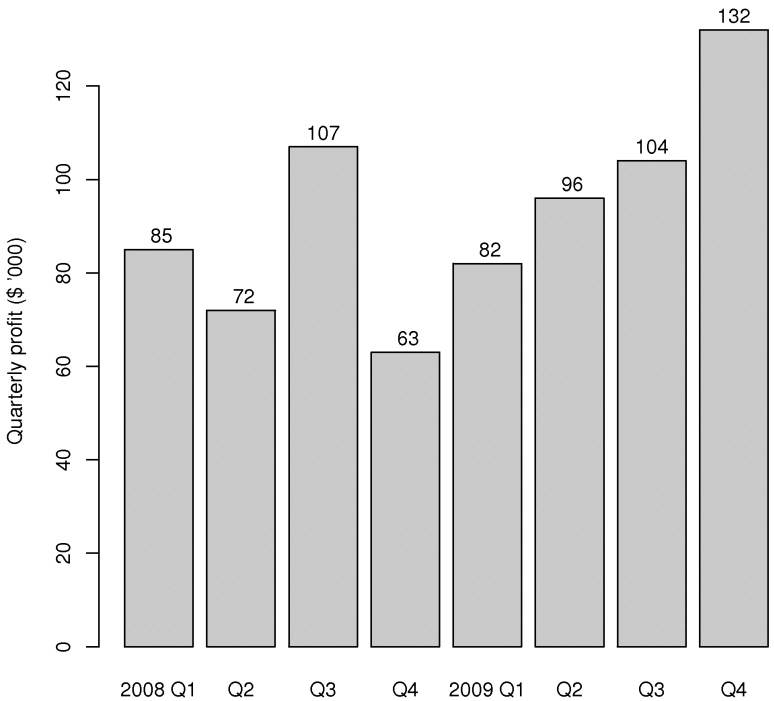


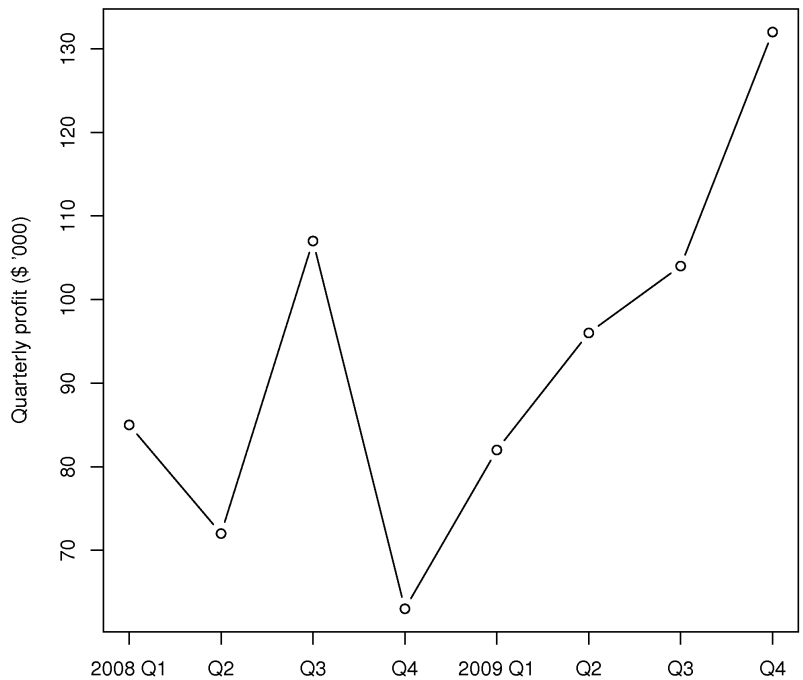
Use a bar plot when:

- many categories
- interpretation does not change if category axis is reordered

Bar plots

Rather use a time-series plot if the data have a sequence:

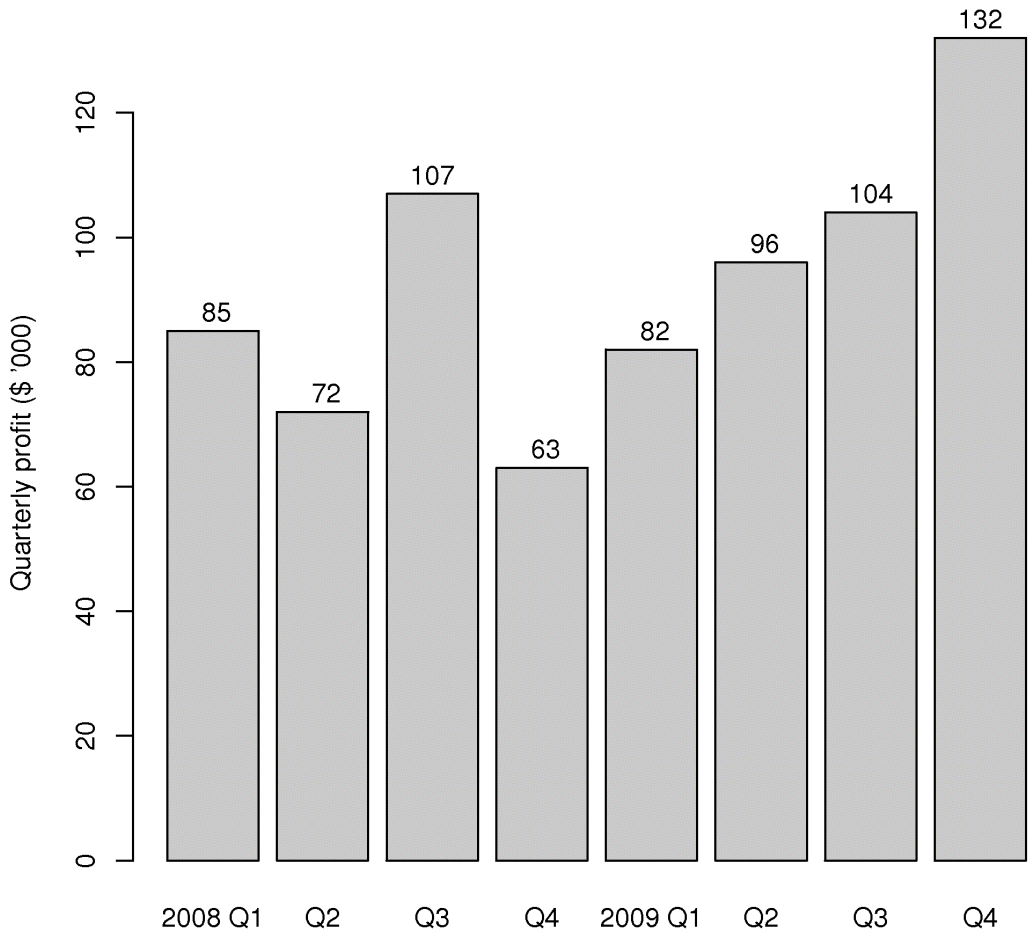




You can see the trends more clearly.

Bar plots

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

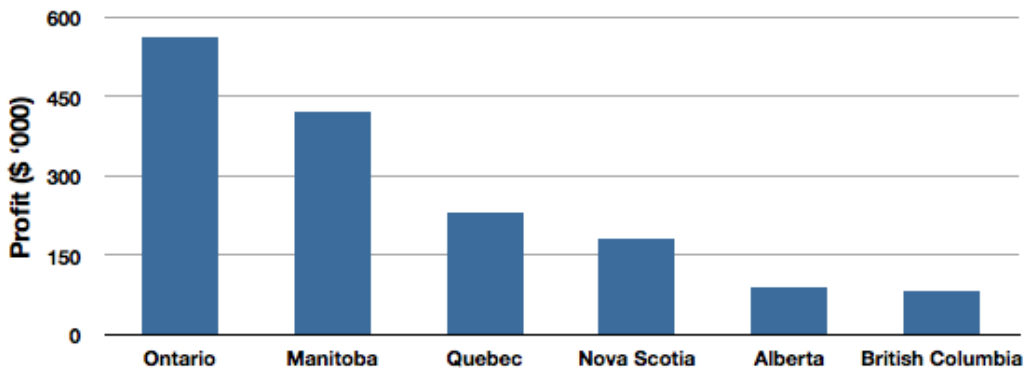
Bar plots

Maximize data ink ratio within reason

$$\text{Maximize data ink ratio} = \frac{\text{total ink for data}}{\text{total ink for graphics}} = 1 - \text{proportion of ink that can be erased without loss of data information}$$

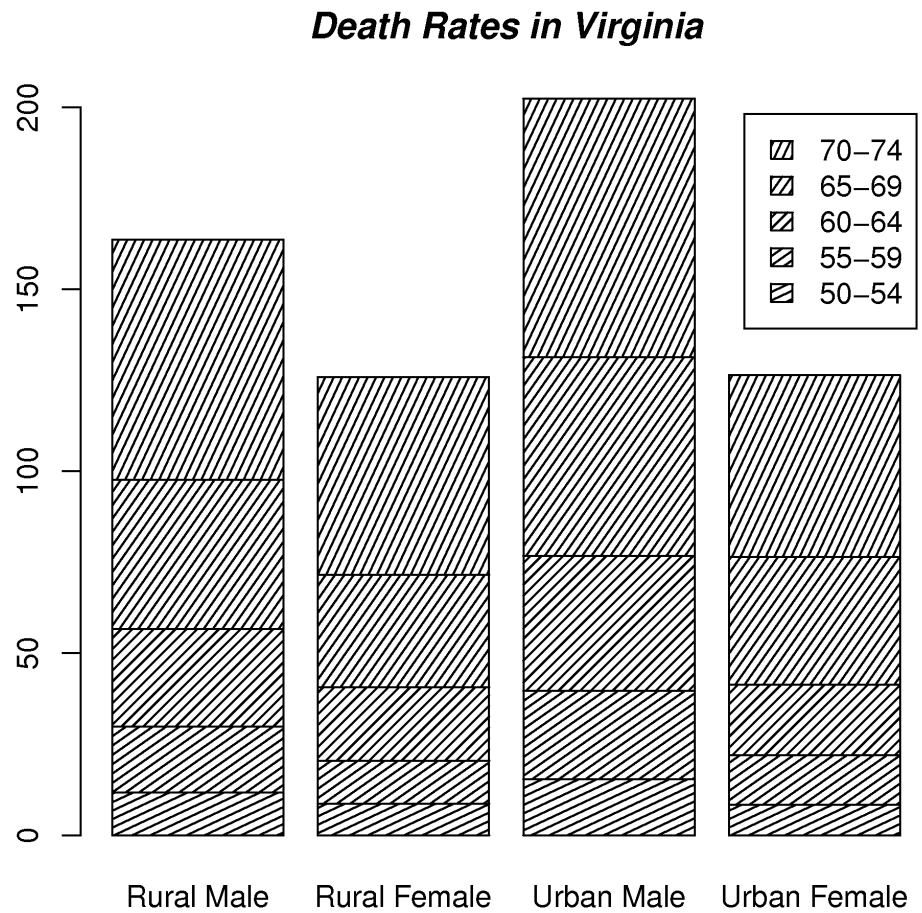
- Rather use a table for a handful of data points:

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

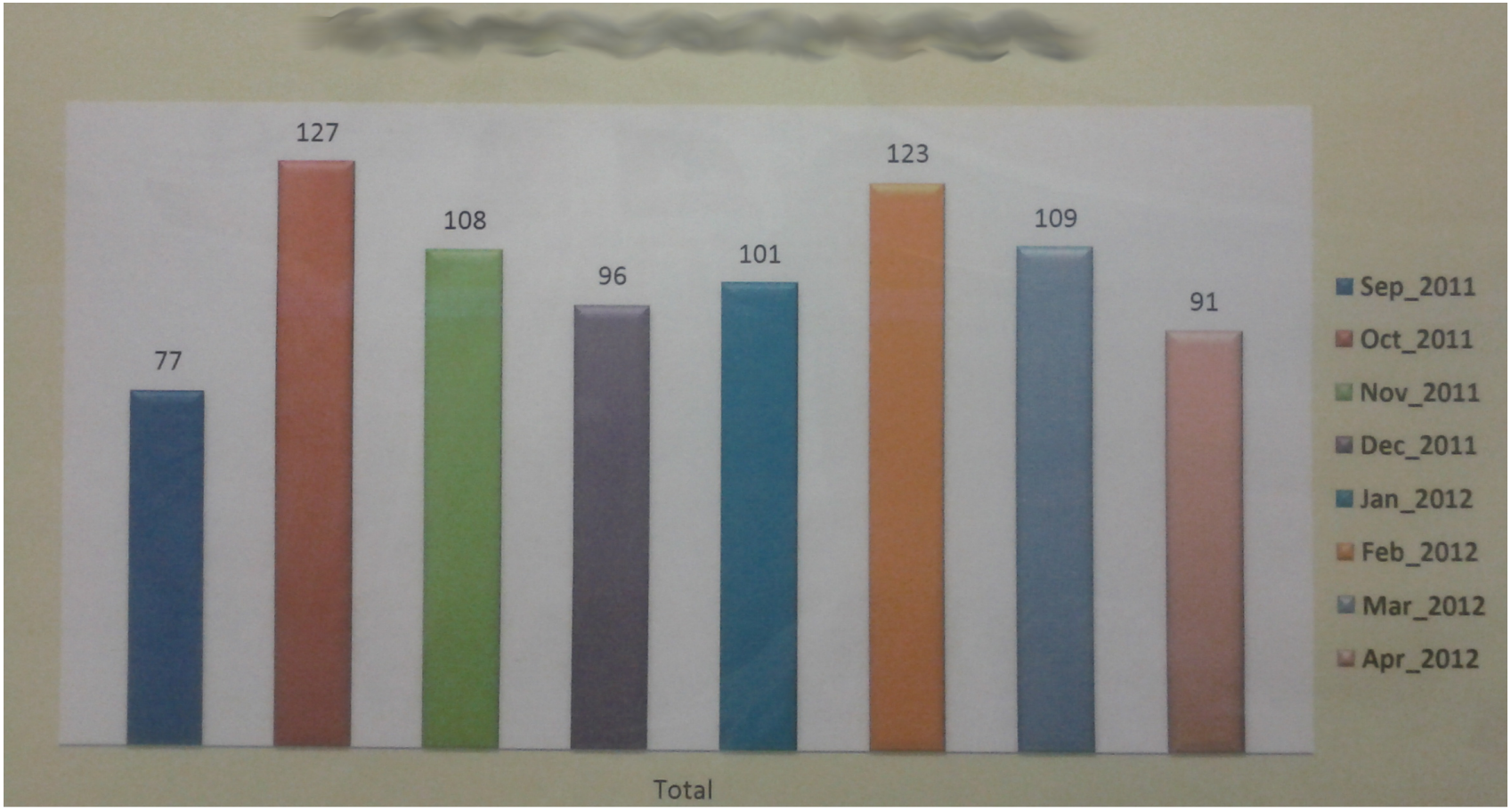


Bar plots

- 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

- Use horizontal bars 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

Box plots

A graphical display of the "5-number summary" for 1 variable

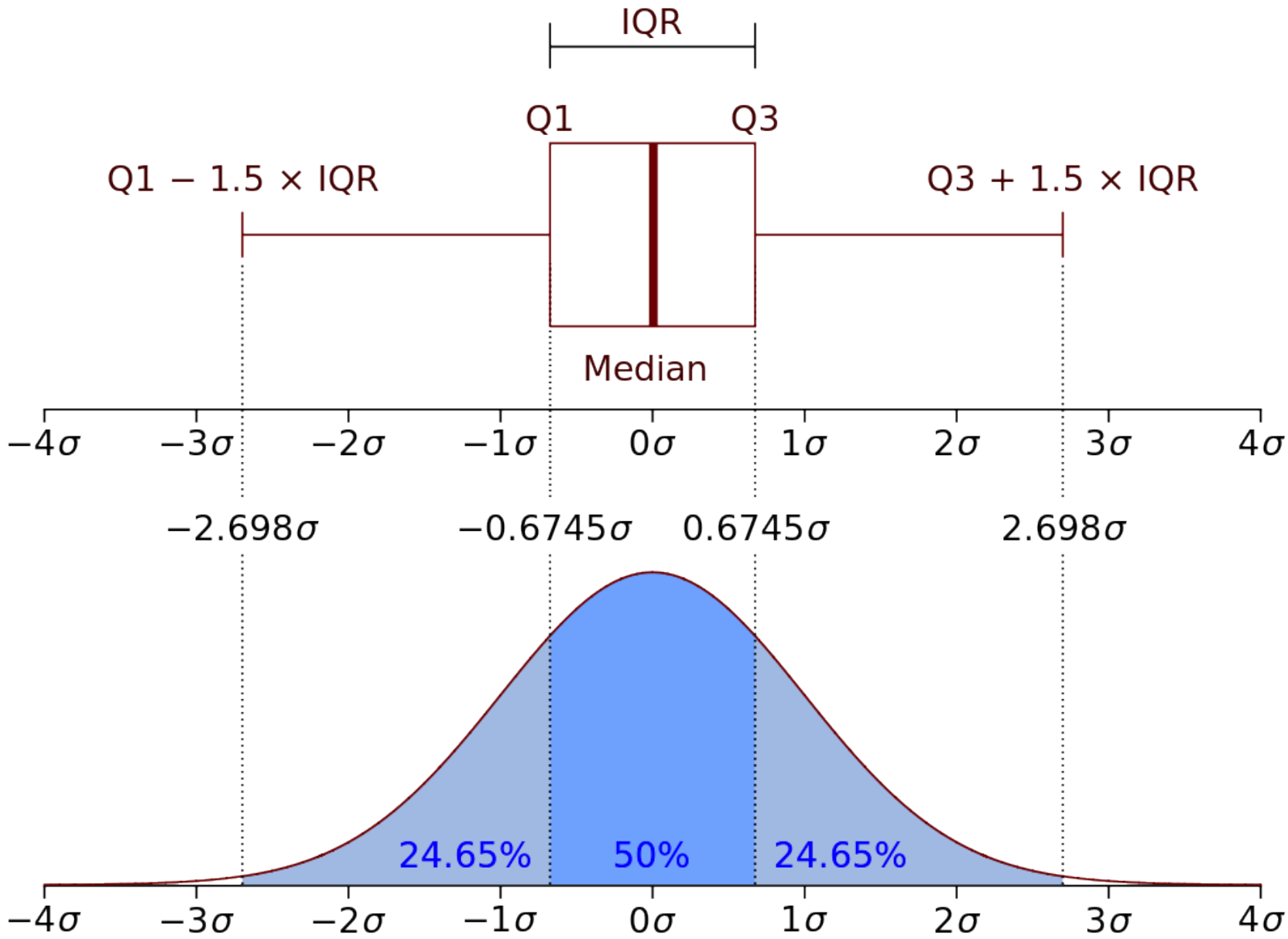
- whisker = minimum sample value [*or*: median $-1.5 \cdot$ IQR]
- 25th percentile (1st quartile)
- 50th percentile (median)
- 75th percentile (3rd quartile)
- whisker = maximum sample value [*or*: median $+1.5 \cdot$ IQR]

Notes:

1. 25th percentile is the value below which 25 percent of the observations in the sample are found
2. distance from 3rd to 1st quartile = interquartile range (IQR)

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

Box plots: compared to a pure normal distribution

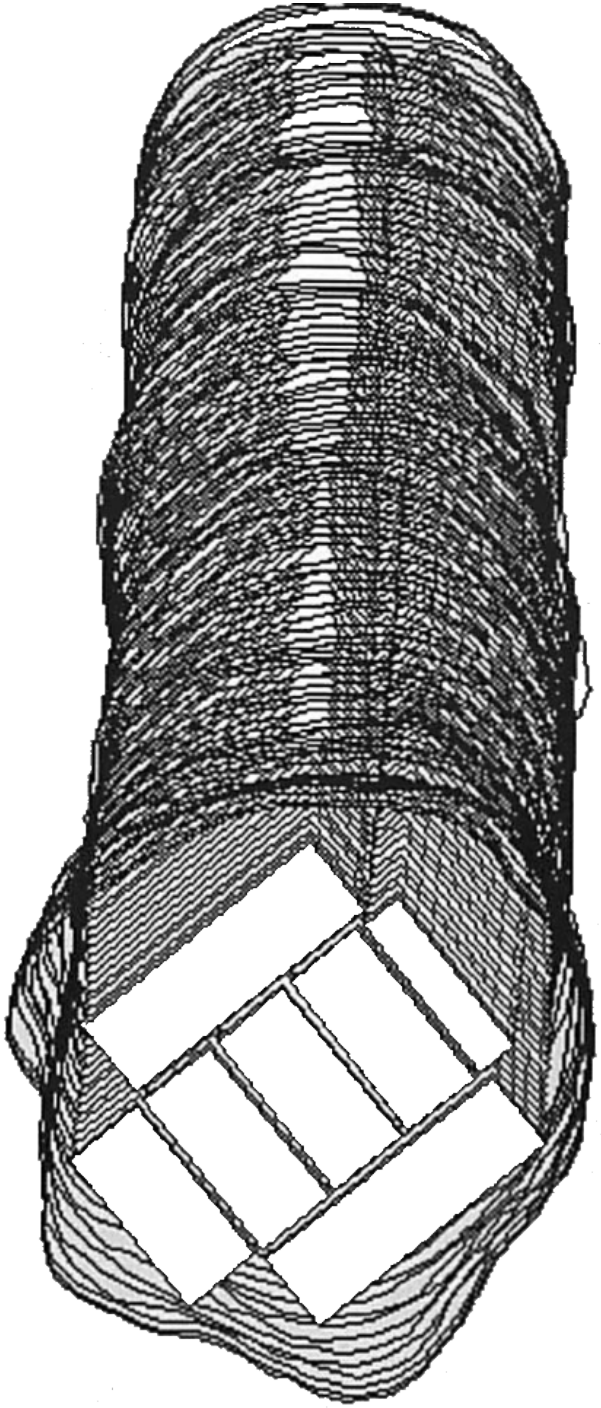


[Wikipedia has some really great illustrations to explain statistical concepts]

Box plots

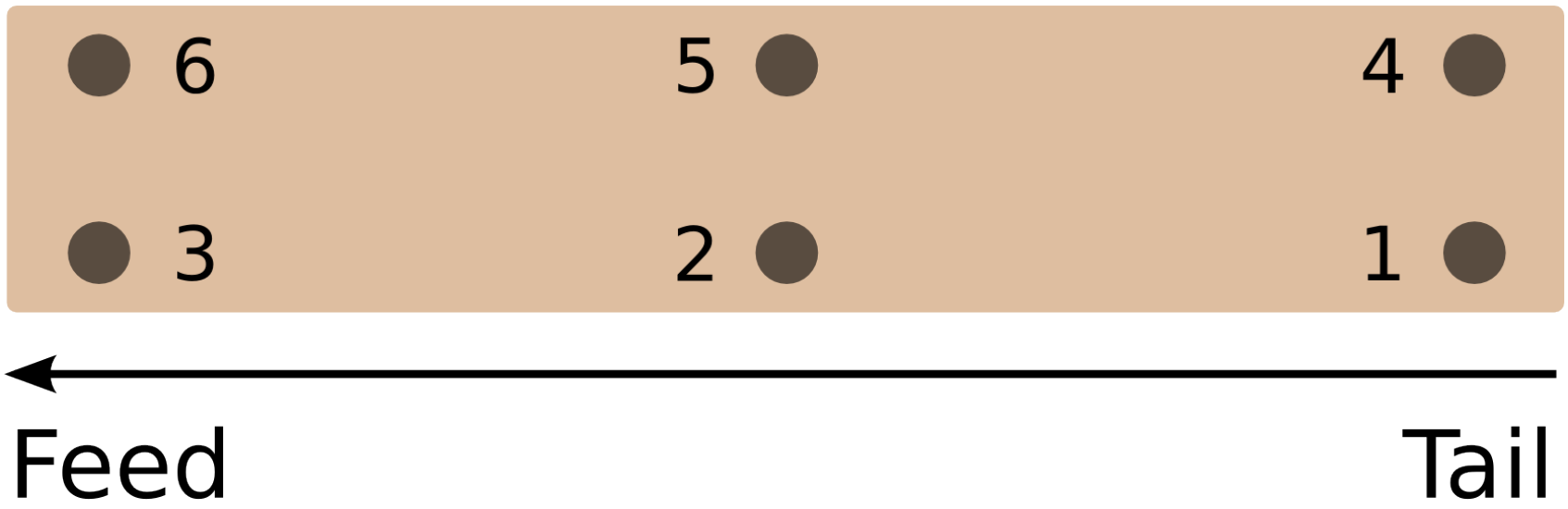
Video of data source: sawmill in Québec

- 4 degrees of rotation of log as it moves through the saws



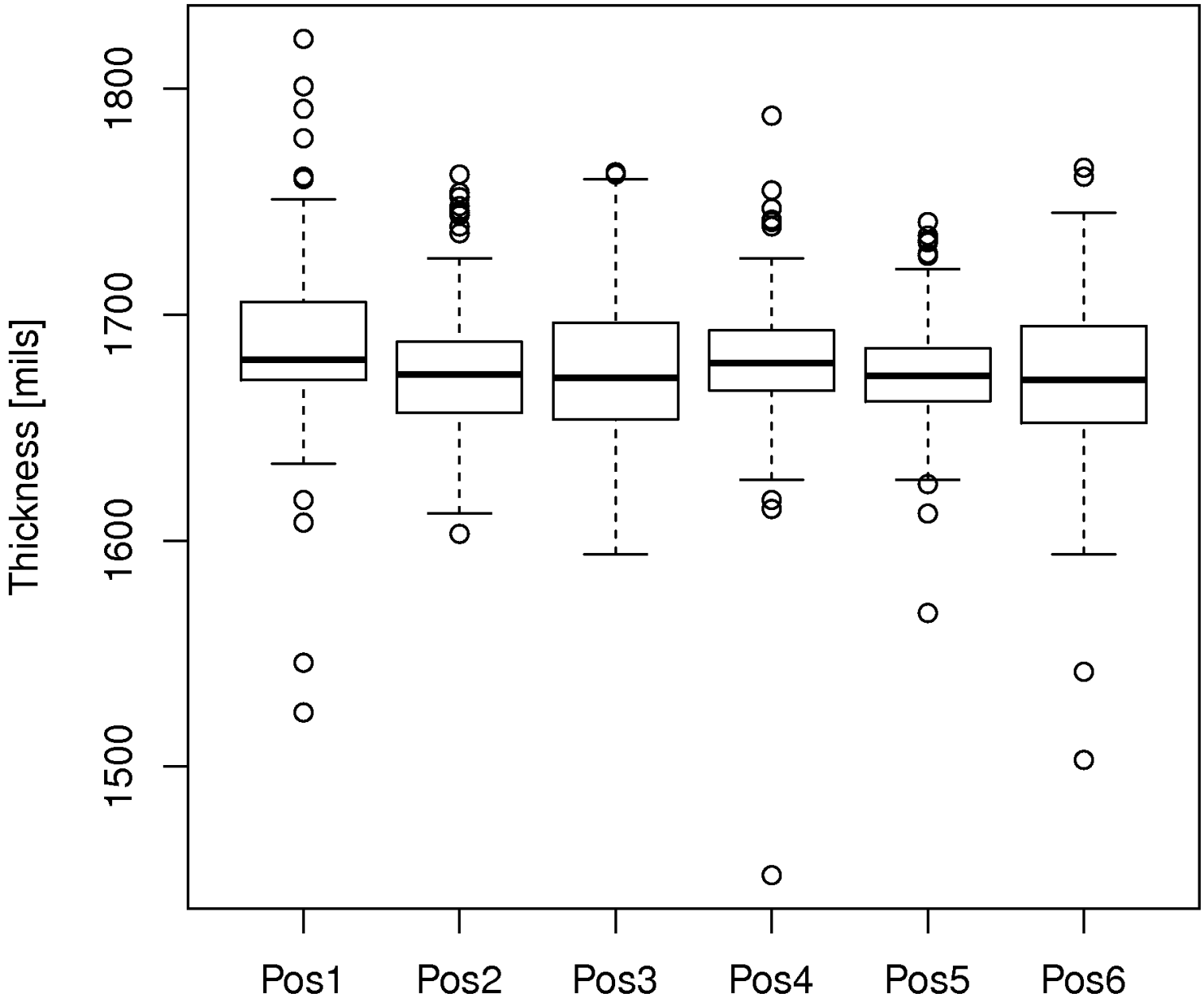
Box plots

Thickness measured at 6 locations; target = 1680 mils



Actual 2x6 thickness = 1500 mils; extra for the lumber to dry out

Box plots

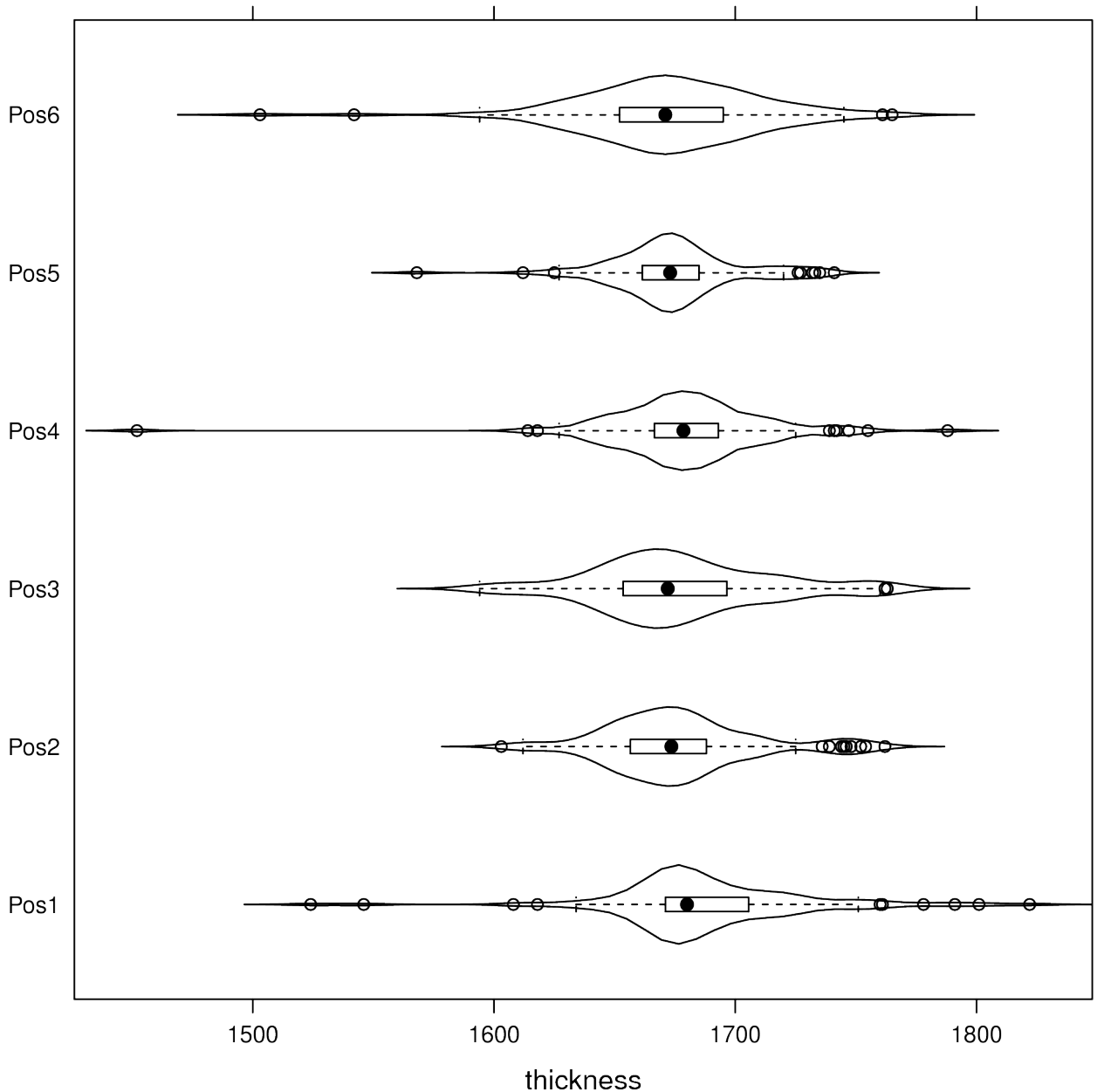

Thickness [mils]
1500
1600
1700
1800
Pos1
Pos2
Pos3
Pos4
Pos5
Pos6

Box plots

Some variations:

- 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)

Box plot variation: violin plot



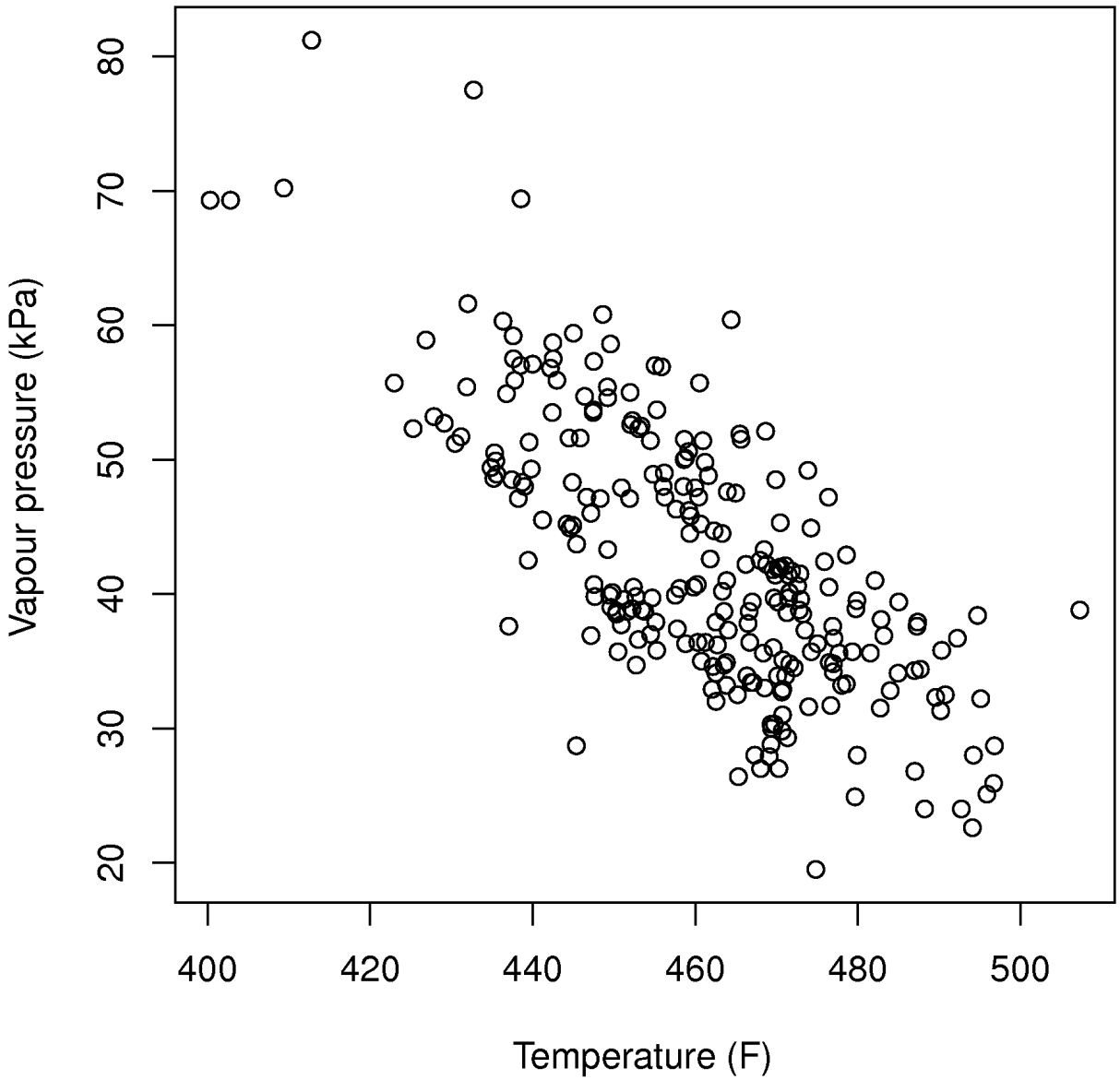
Scatter plots

- Used to help understand the relationship between two variables: a bivariate plot
- Collection of points in the 2 axes
- Each point is the intersection of the values on each axis

Intention of a scatter plot

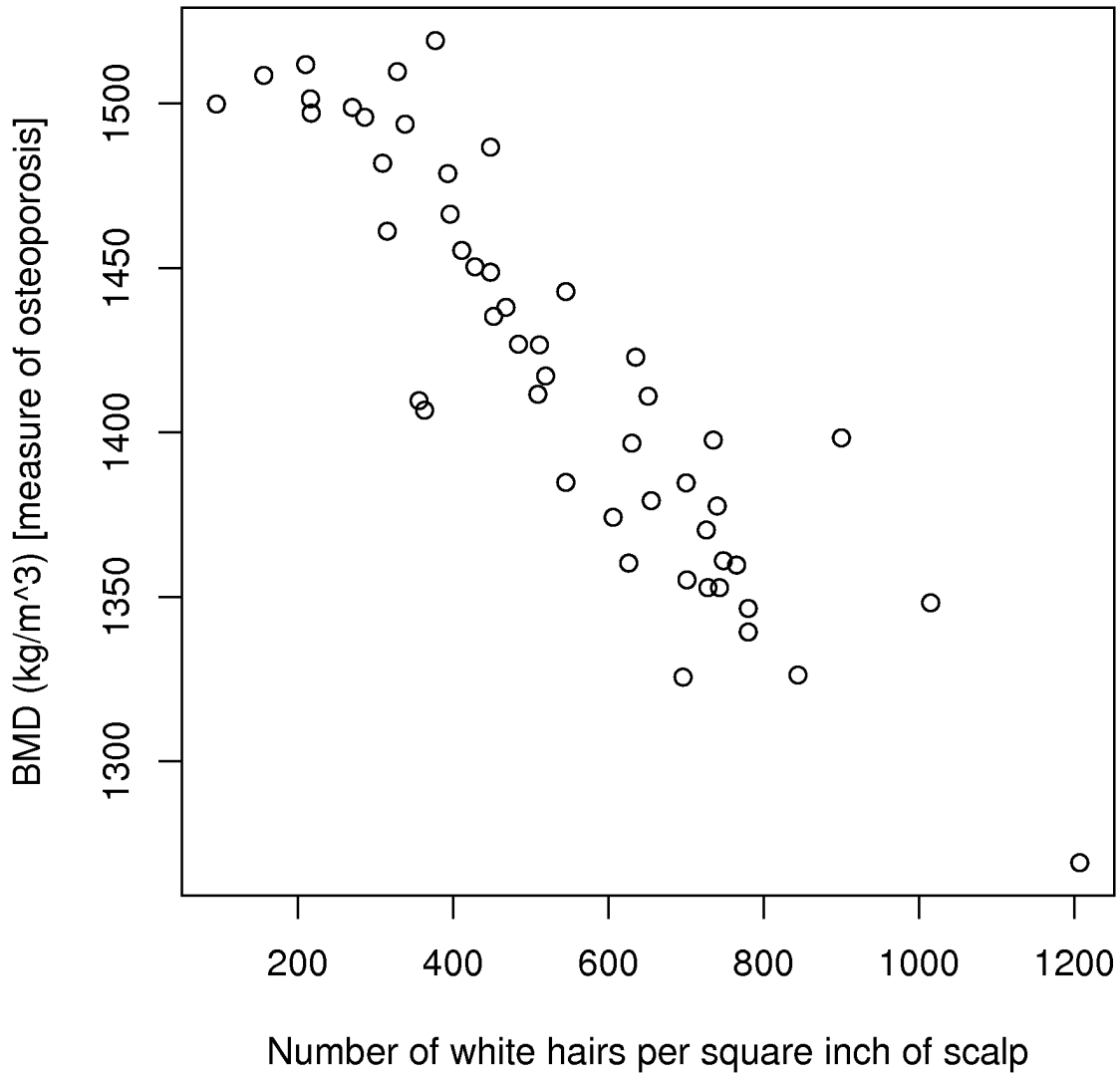
Asks the viewer to draw a causal relationship between the two variables

Scatter plots



Scatter plots

However, not all scatter plots show causal phenomenon.



Scatter plots

Strive for graphical excellence by:

- making each axis as tight as possible
- avoid heavy grid lines
- use the least amount of ink
- do not distort the axes

Scatter plots

There is an unfounded fear that others won't understand your 2D scatter plot.

- Tufte study (VDQI): no scatter plots in a sample (1974 to 1980) of Western dailies
- 12 year olds can interpret such plots.
- Japanese newspapers frequently use scatterplots
- Plant control room: seldom see scatter 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.

Here's an example (January 2013 publication)

- Why did the author use a time-series plot to show correlation?
- Would the plot be more informative as 2D-scatter plot?
- What if you were to repeat this analysis for multiple regions/-countries/cities. How would you show (visualize) the correlations effectively?

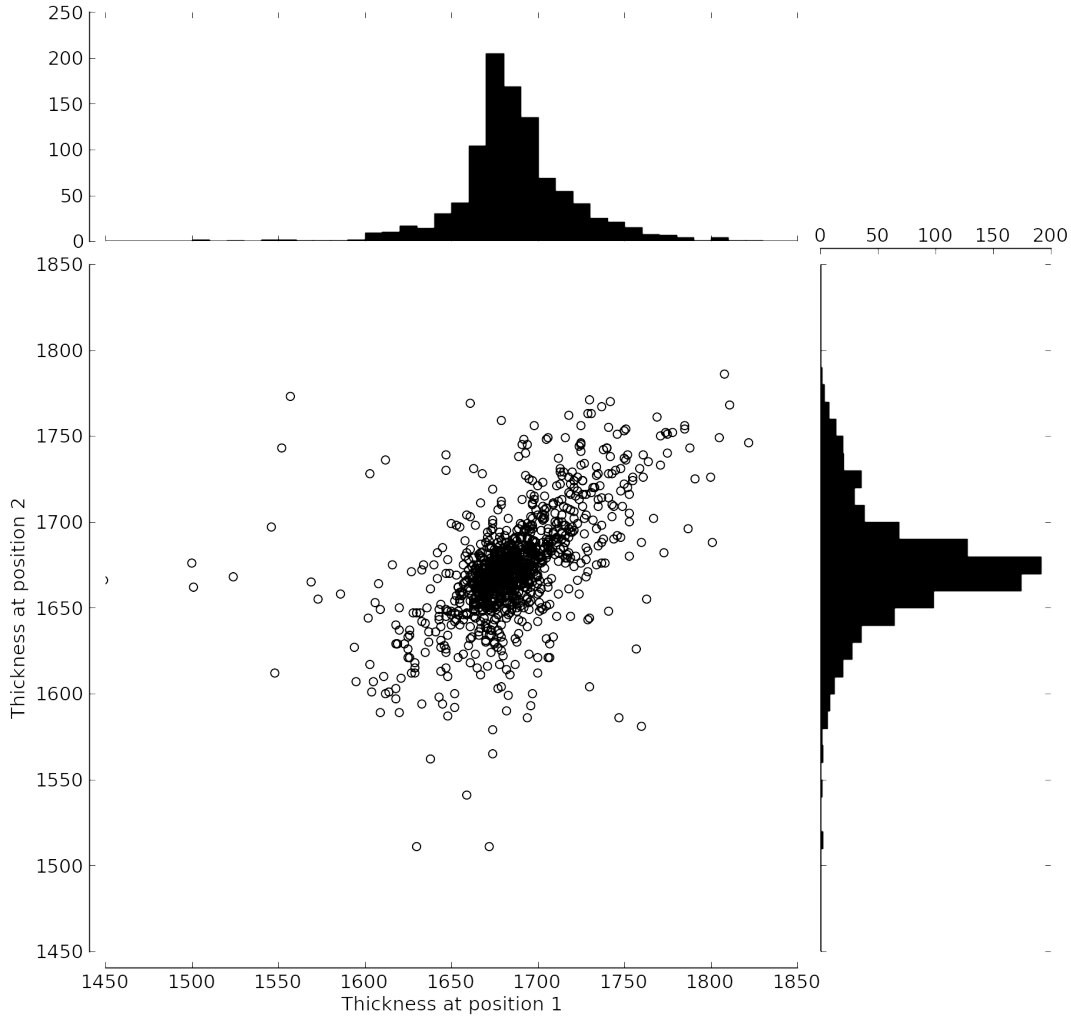
$Pb(CH_2CH_3)_4$

[Read the full story for more interesting details and geographic visualizations:

http://www.motherjones.com/environment/2013/01/lead-crime-link-gasoline]

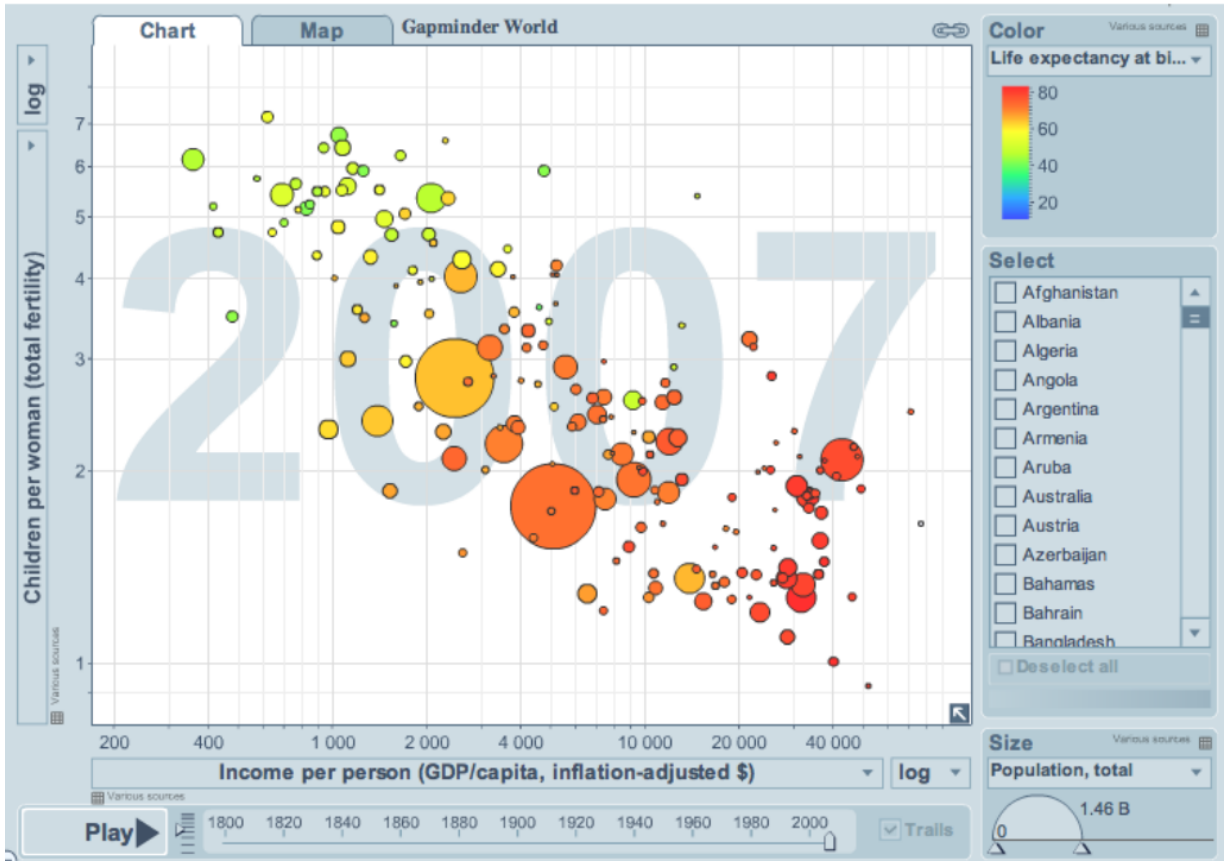
Scatter plots

- Add box plots or histograms to assist interpretation:



Scatter plots

- Add a 3rd variable: different marker sizes
- Add a 4th variable: use colour or grayscale shading



- The GapMinder website allows you to "play" the graph over time (the 5th variable)

Scatter plots

- Web-based demo from http://gapminder.org
- Demo by Hans Rosling (requires internet access)

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?

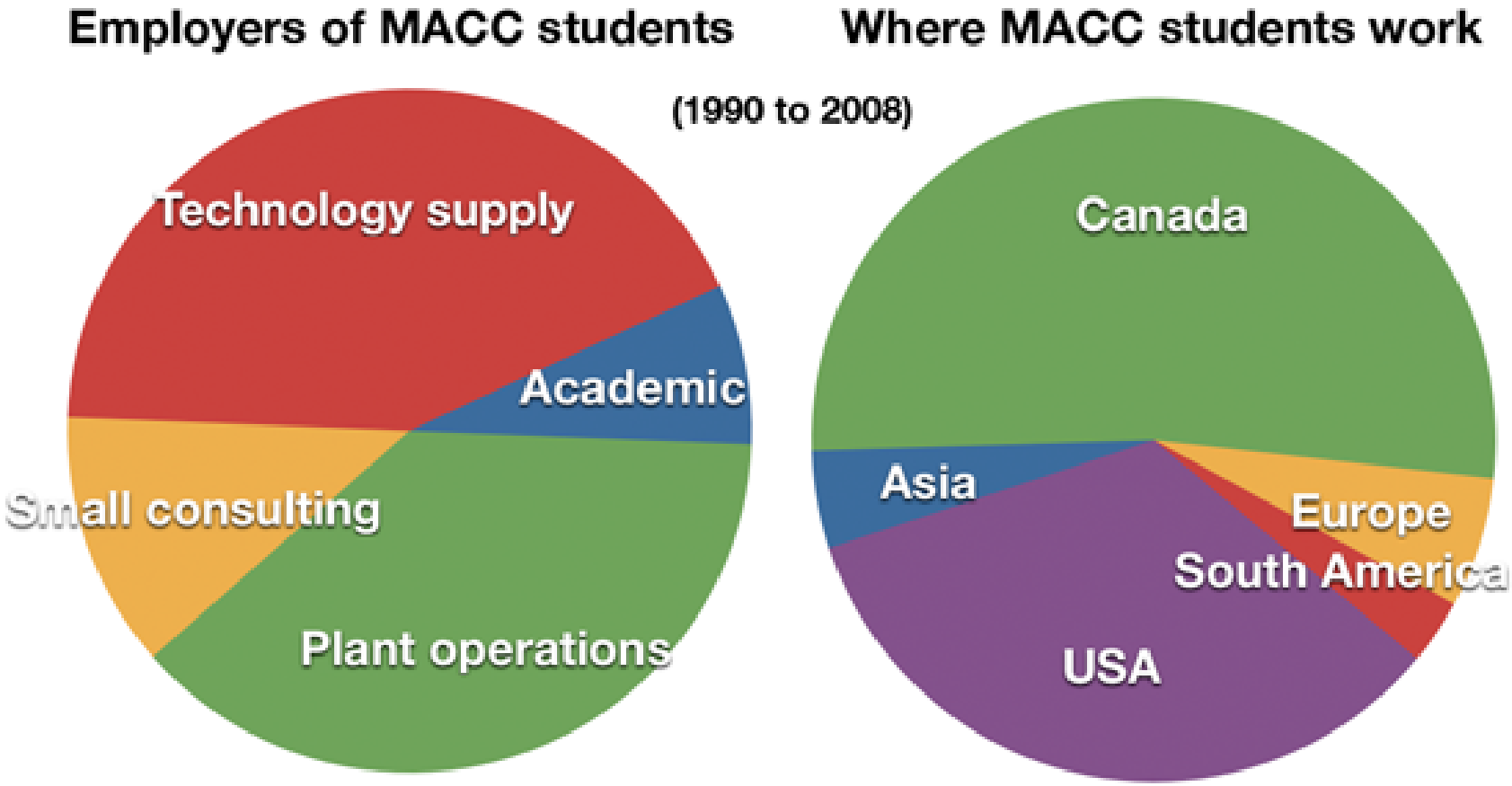
Tables

- 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

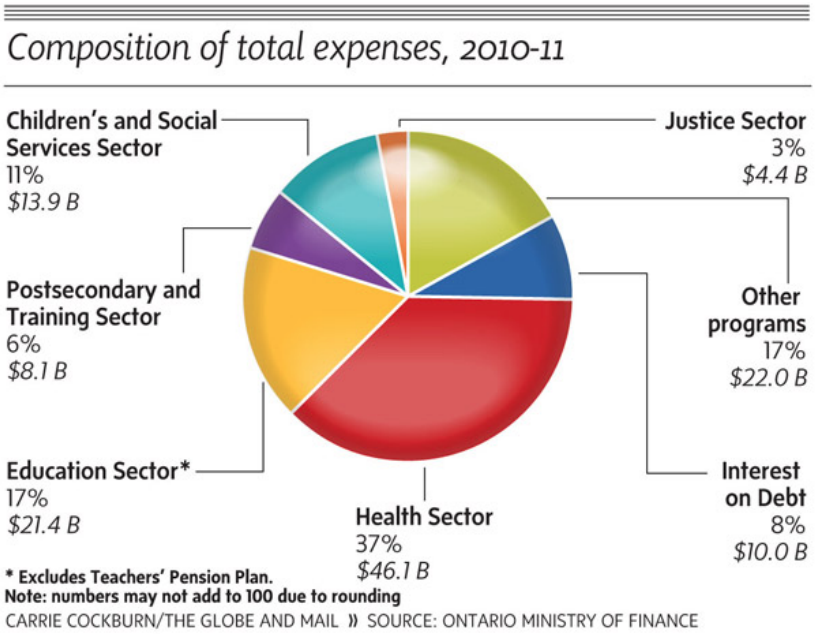
Three common pitfalls:

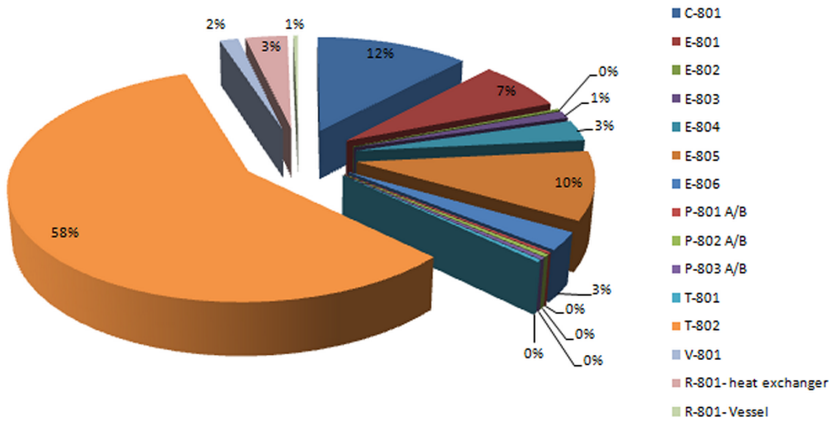
1. using pie charts when tables will do

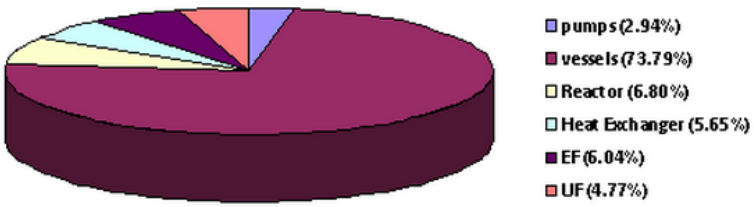


I cannot explain the pitfalls of pie charts as well as Stephen Few does: Save the pies for dessert (please read)

Tables vs pie charts: plenty of bad examples







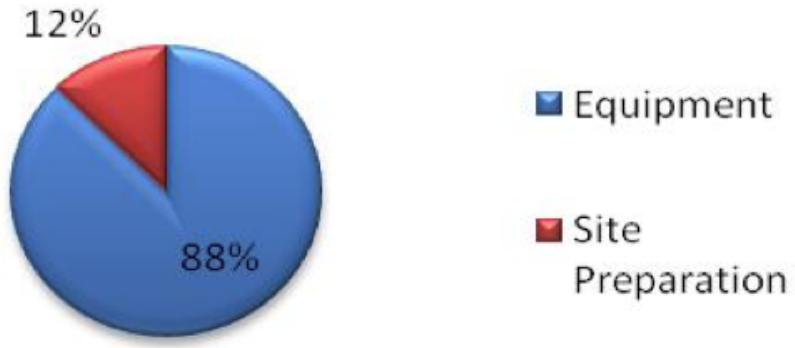


Figure 2 Total capital cost distribution among equipment

[Globe and Mail, March 2010 (top left); SDL reports, 4N4, 2012 (all others)]

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**

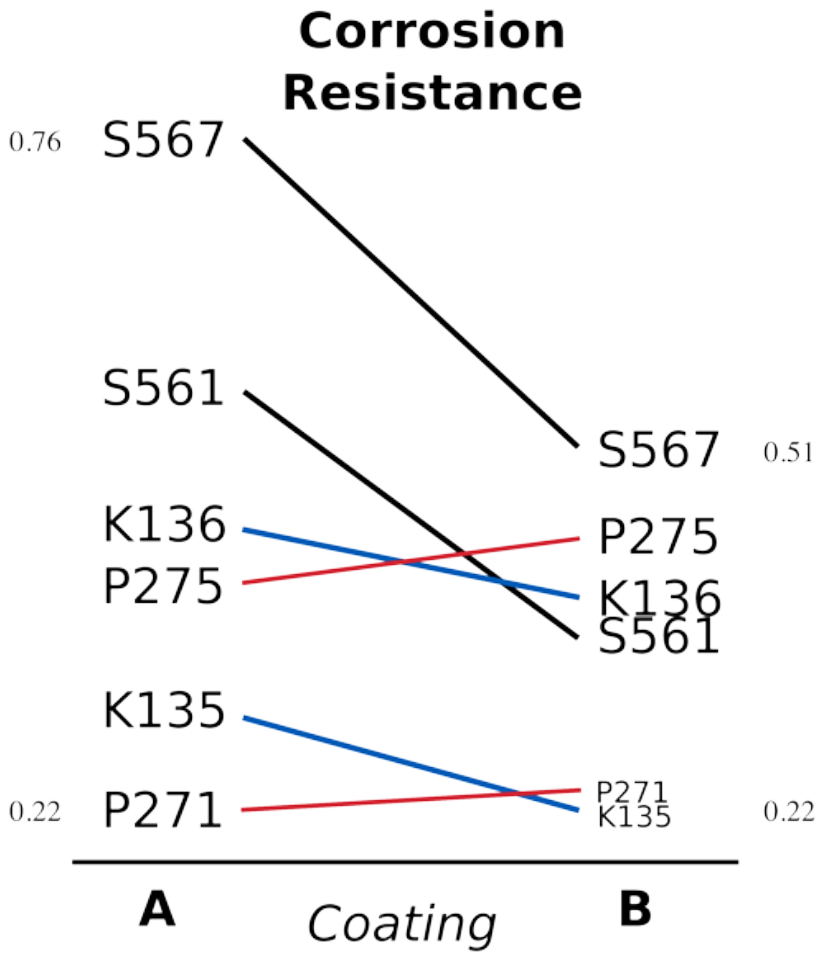
Tables

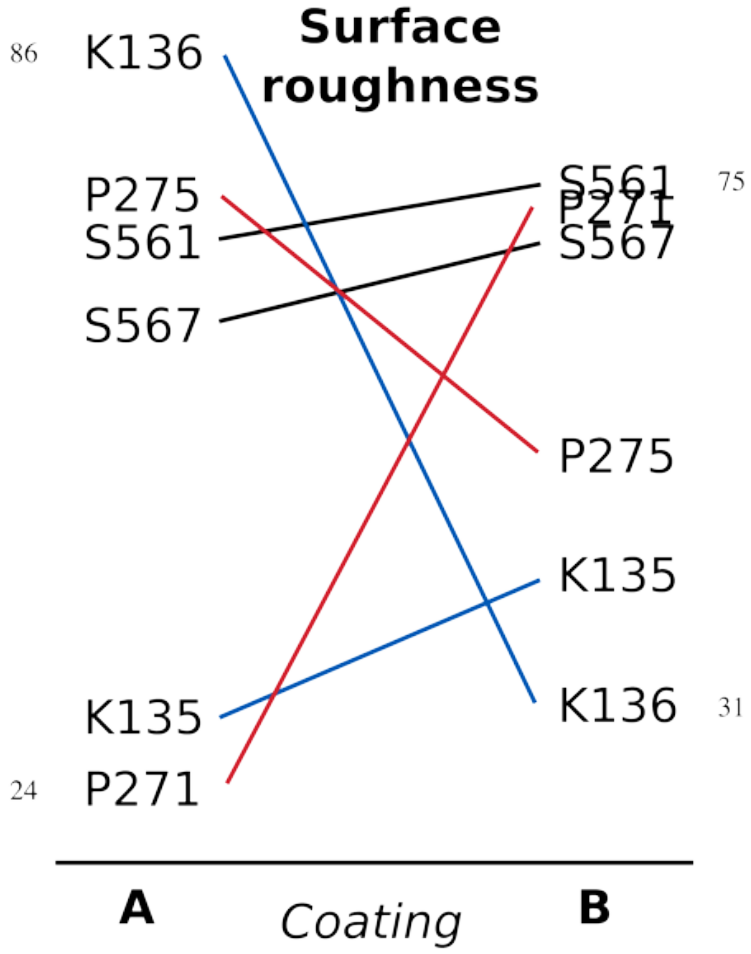
Interesting example: comparing two treatments

	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
P271
0.22
S567
0.51
P275
K136
S561
P271
K135
0.22
A
Coating
B


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

Data frames

Frames are the basic containers that surround the data and give context to our numbers. Here are some tips:

1. Use round numbers
2. Tighten the axes as much as possible, except ...
3. when showing comparison plots: *all axes must have the same minima and maxima*

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:

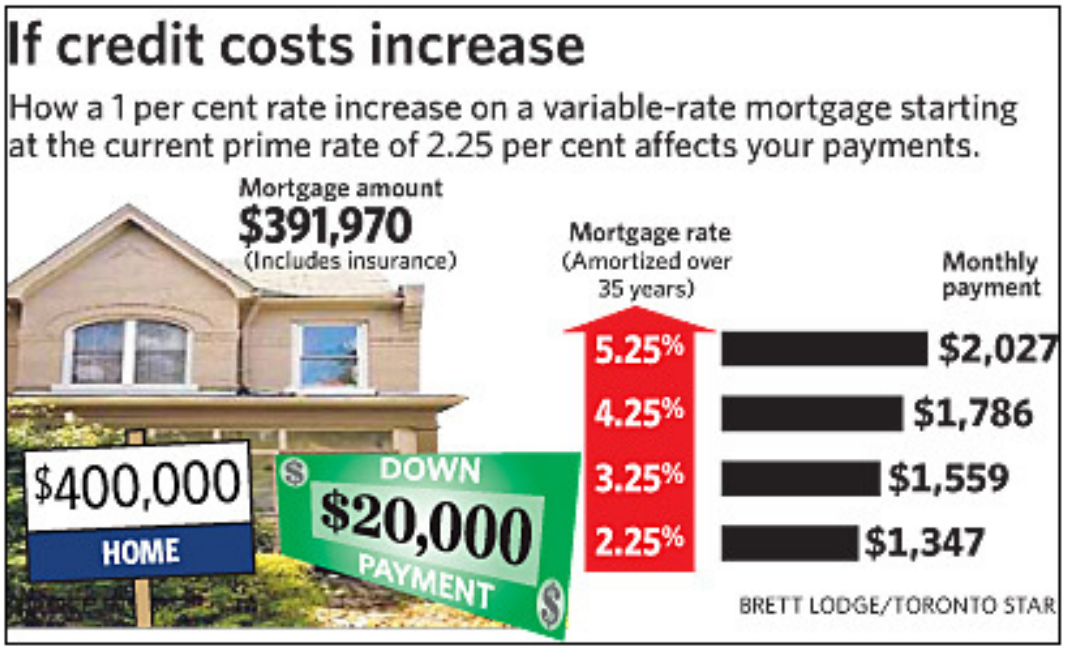
- To understand causality, you must show causality: use bivariate scatter plots (sometimes line plots also work well)
- Plots and text go together: a plot = paragraph of text
  - add labels to plots for outliers and interesting points
  - add equations
  - add small summary tables
- Avoid codes: "A =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