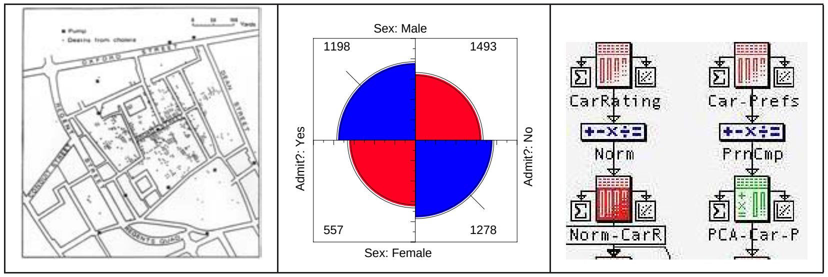
The Past, Present and Future of Statistical Graphics

(An Ideo-Graphic and Idiosyncratic View)



Michael Friendly

York University

http://www.math.yorku.ca/SCS/friendly.html

VIEWS, London, Nov, 2004

Tables and graphs: Tasks, goals, audience

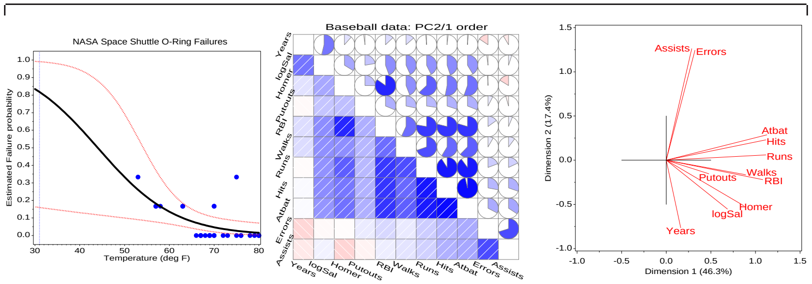
Like good writing, effective graphical displays require an understanding of purpose—what is to be communicated, and to whom Friendly (1991)

- **Tasks and Goals for information display**
  - Lookup— read off exact numbers
  - Comparisons— which is more?
  - Detecting patterns, trends, anomalies
  - Different tables or graphs for different purposes: analysis, persuasion
  - Visual presentation as *communication*:
    - what do you want to say?
    - what the the audience?
- **Tables vs. Graphs**
  - Tables are best suited for *look-up*— read off exact numbers
  - Graphs are better for showing *patterns, trends, anomalies*, making *comparisons*

Part 2: Tables and Graphs: Some principles of Graphical Display

If I can't picture it, I can't understand it Albert Einstein

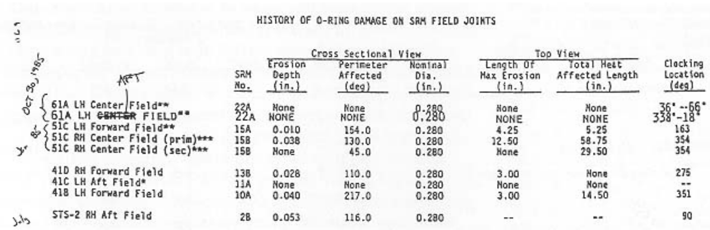
- Graphical failures and successes
- Graphical comparisons and graphical perception
- Corrgrams: rendering and variable order
- Effect ordering for data display



Graphical failure: *Challenger* disaster

What we have here is a failure to communicate Cool Hand Luke

- Few events in history provide as compelling an illustration of importance of appropriate ordering and display of information.
- Tables and charts presented to NASA by Thiokol engineers showed data from prior launches ordered by *time* (launch number), rather than by *temperature*— the crucial factor.

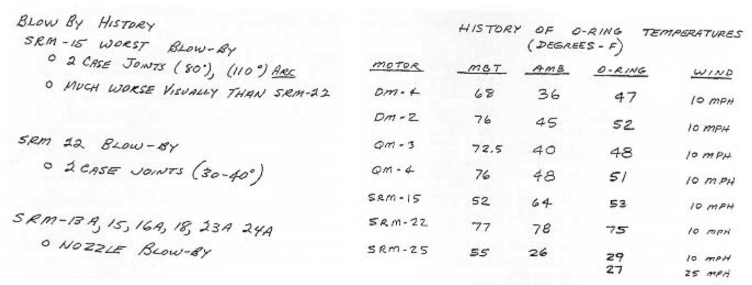


HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

		Cross Sectional View			Top View		
	SRM No.	Erosion Depth (in.)	Perimeter Affected (deg)	Nominal Dia. (in.)	Length Of Max Erosion (in.)	Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36°-66°
61A LH ~~CENTER~~ FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

Graphical failure: *Challenger* disaster

- The engineers' charts were also remarkable for *information muddeling*— extraneous information (wind), cryptically abbreviated labels, no clear assessment of damage ("blow-by" (soot) vs. "erosion depth" (O-ring damage)).



BLOW BY HISTORY
SRM-15 WORST BLOW-BY
o 2 CASE JOINTS (80°), (110°) ARC
o MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY
o 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A
o NOZZLE BLOW-BY

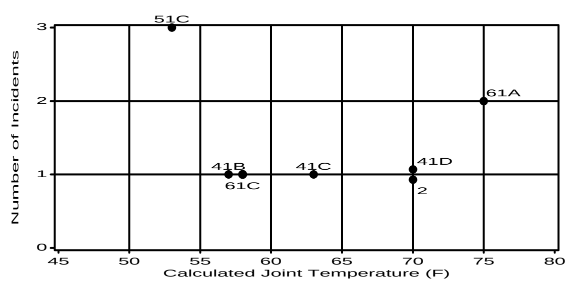
HISTORY OF O-RING TEMPERATURES (DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
DM-4	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

- Engineers *did* make the proper recommendatation: "O-ring temperature must be $\geq 53°$F at launch." NASA launch control over-rode the recommentation.

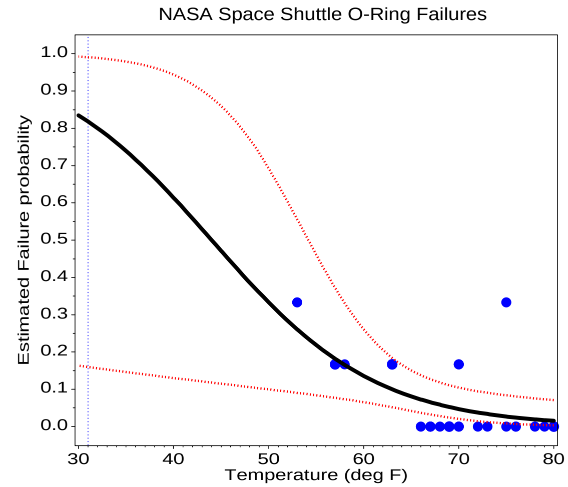
Graphical failure: *Challenger* disaster

- Tufte (1997) notes:
  - "the fatal flaw is in the *ordering* of the data,"
  - "the graphics... suggest there are right ways and wrong ways to display data; there are displays that reveal the truth and displays that do not."
- Thiokol engineers did prepare a graph— but it was seriously misleading. (What are the flaws?)



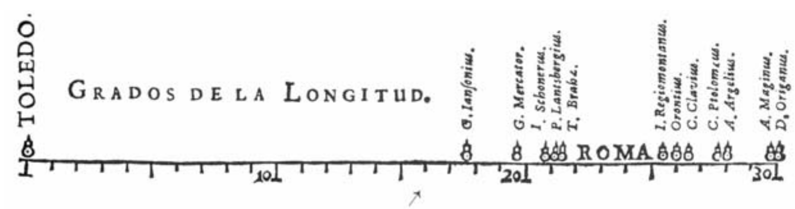
Graphical failure: *Challenger* disaster

- A better display shows all the data, some prediction, and an an indication of uncertainty. It is hard to imagine a launch at $31°$F given this graph.



Graphical success: van Langren's graph of longitude

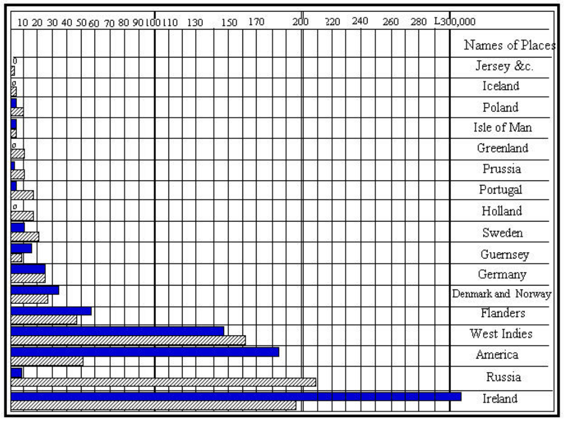
- van Langren could have presented these data as a table— sorted by date (priority), name (provenance), or value (range)
- Only his hand-drawn graph shows simultaneously:
  - individual estimates and spacings along the scale
  - associated names, offset to avoid overlap
  - estimated, central value ('ROMA') and wide variability



Graphical success: Playfair's first barchart

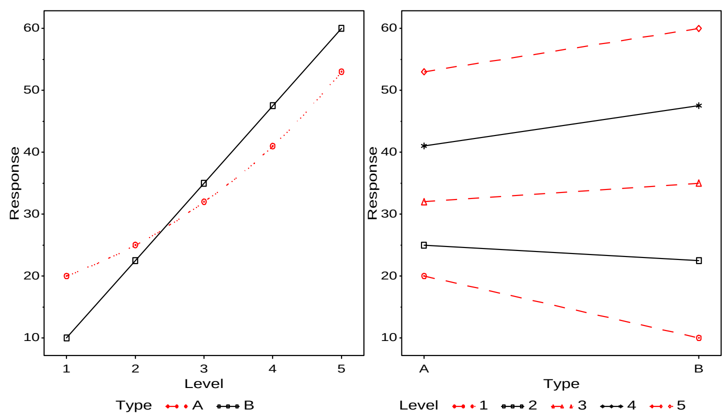
Imports and exports of Scotland (Playfair, 1786)

- Horizontal, to show the "country" labels
- Grouped by country, so imports/exports could be directly compared.
- Sorted by numerical value rather alphabetically by country (as would be done by most statistical graphing software)



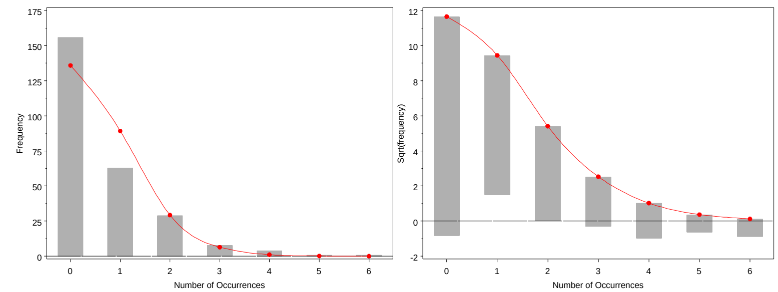
Graphical comparisons: Make them easy

- Visual grouping— connect with lines, make key comparisons contiguous
  - Left: easier to compare across Level
  - Right: easier to compare across Type



Graphical comparisons: Baselines

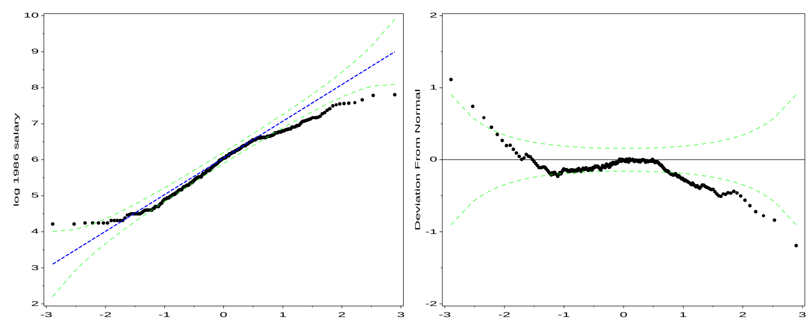
- Baselines— compare *data* to *model* against a line, preferably horizontal
  - Comparing observed and fitted discrete distributions: histogram and hanging histogram



See: `http://www.math.yorku.ca/SCS/vcd/rootgram.html` for hanging histograms and hanging rootograms.

Graphical comparisons: Tolerances

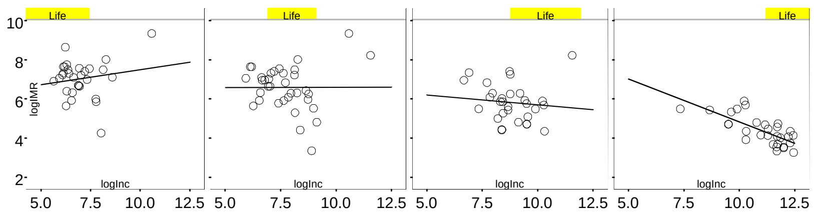
- Tolerances— show an acceptable region around a comparison standard
  - Normal QQ plot: Standard vs. Detrended



See: `http://www.math.yorku.ca/SCS/sssg/nqplot.html`

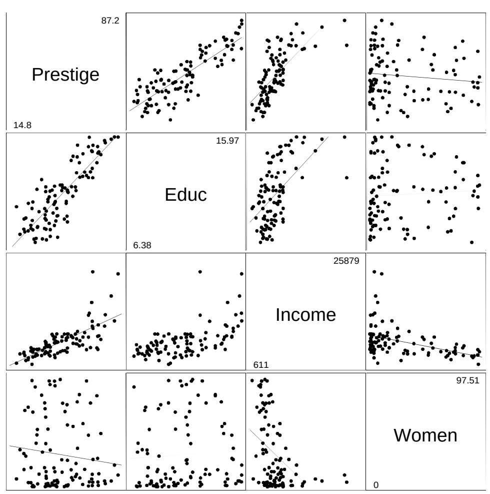
Graphical comparisons: Small multiples

- Multiple, contiguous panels allow differences to be sensitively compared
- e.g., Coplots of log(Infant Mortality) vs. log(Income) | Life Expectancy



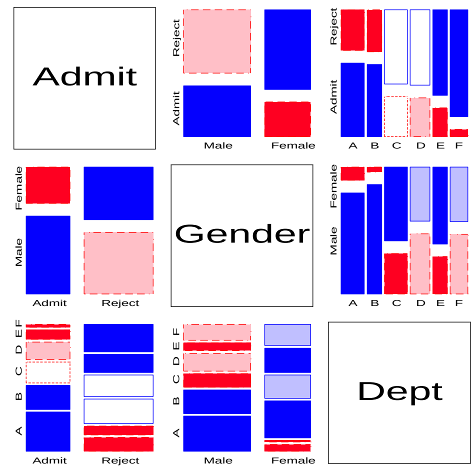
See: `http://www.math.yorku.ca/SCS/sasmac/coplot.html`

- e.g., scatterplot matrix for quantitative data: all pairwise scatterplots



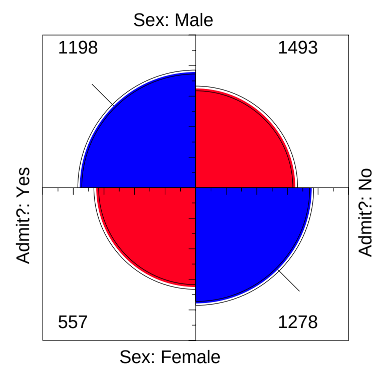
Graphical comparisons: Small multiples

- e.g., mosaic matrix for quantitative data: all pairwise mosaic plots

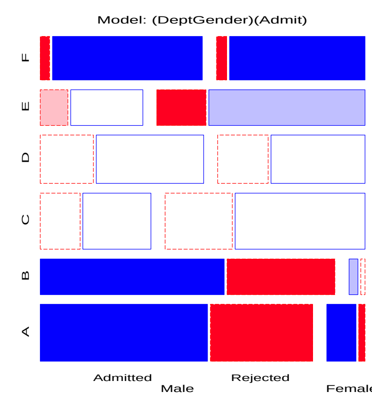


Visual codes for Quantative vs. Frequency data

- Quantitative data: **magnitude** $\sim$ **position along an axis**
- Frequency data (Friendly, 1995): **count** $\sim$ **area**



Fourfold display for 2×2 table

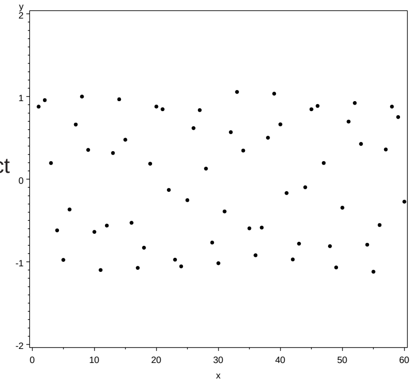


Mosaic plot for 3-way table

Graphical comparisons: Aspect ratios

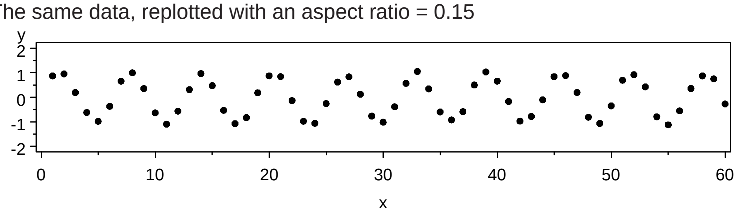
- Shape of a plot (height/width)— *aspect ratio*— often determines what you can see.
- Typically chosen by software to fill the graphics device (landscape, portrait)

- E.g., plot with a square frame (aspect ratio=1)
- Is there any evident pattern here?



Graphical comparisons: Aspect ratios

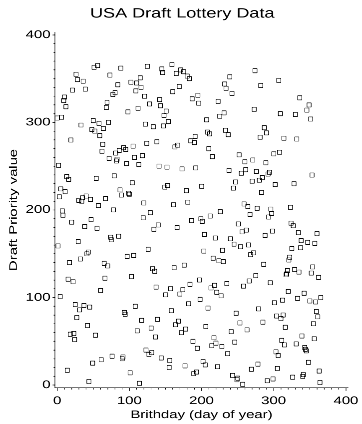
- The same data, replotted with an aspect ratio = 0.15



- General rule: Choose the aspect ratio so the slopes of connecting lines $\approx 45^\circ$.

Smoothing often helps

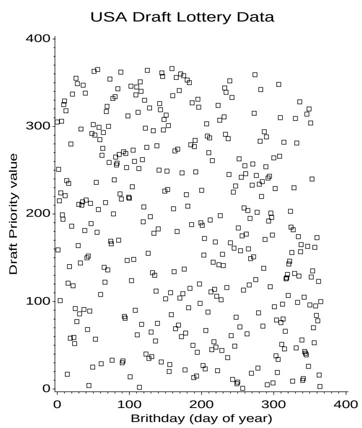
- Our eyes can usually see patterns not easily captured in numbers.
- Sometimes relationships may be too weak to see the trend in a scatterplot.
- Drawing a smoothed curve helps show the trend.

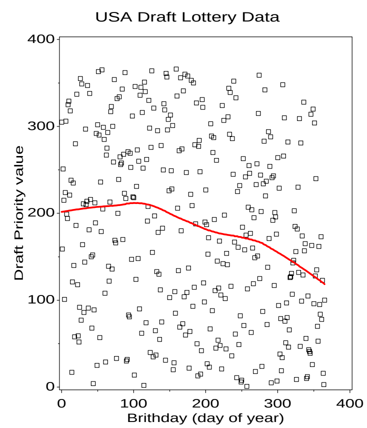


Can you see the trend?

Smoothing often helps

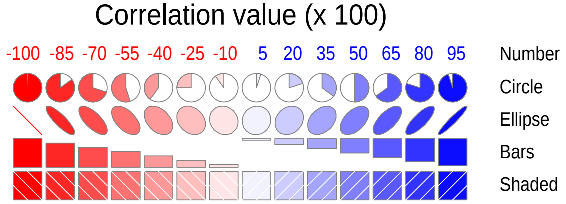
- Our eyes can usually see patterns not easily captured in numbers.
- Sometimes relationships may be too weak to see the trend in a scatterplot.
- Drawing a smoothed curve helps show the trend.





Corrgrams— Correlation matrix displays

- How to show a correlation matrix for different purposes? (Friendly, 2002)
- Render a correlation to depict sign and magnitude (tasks: lookup, comparison, detection)

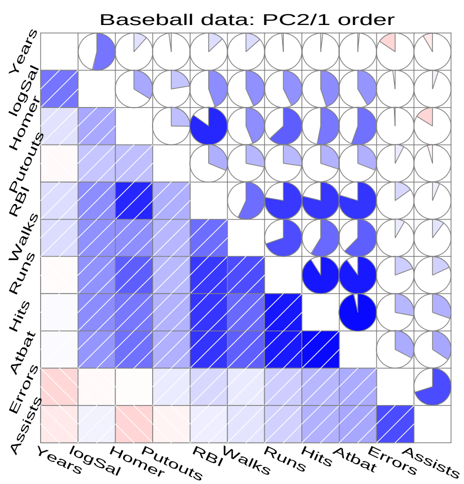


Task-specific renderings:

Task	Lookup	Comparison	Detection
Rendering	Number	Circle	Shading

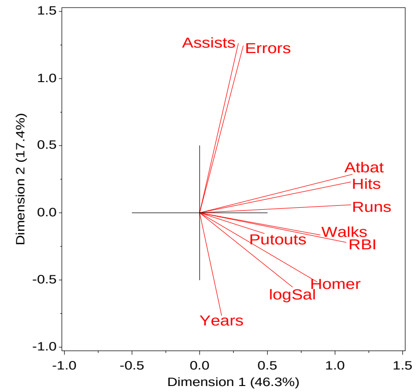
Corrgrams— Rendering

Baseball data: (lower) Patterns vs. (upper) comparison



Corrgrams— Variable ordering

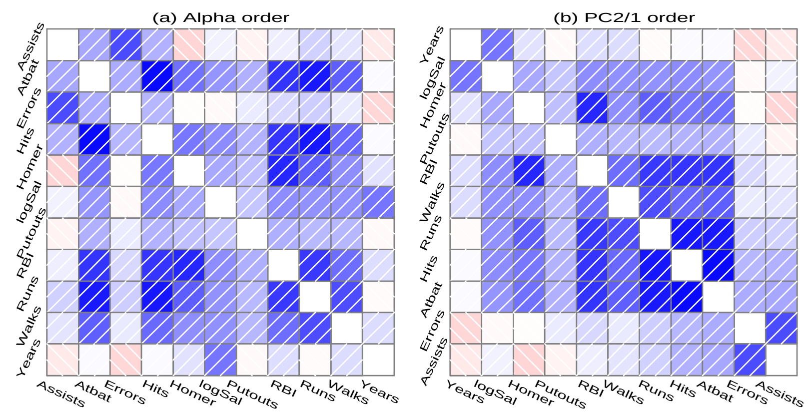
- Reorder variables to show similarities: PC1 or angles (PC2/PC1)



- Generalizations to partial ($R(\boldsymbol{Y} \mid \boldsymbol{X})$), conditional correlations ($r_{ij \mid \text{rest}} \sim \boldsymbol{R}^{-1}$)

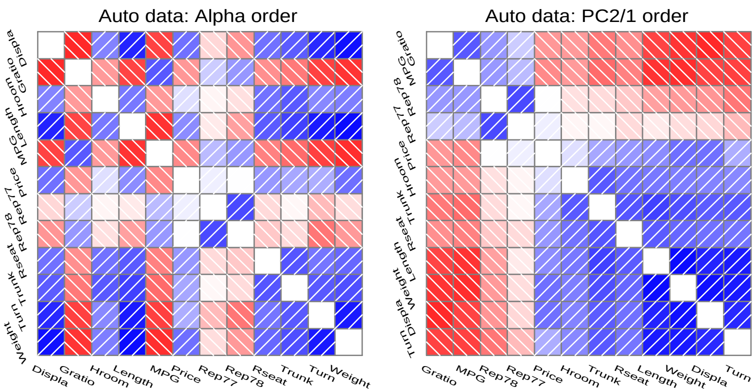
Corrgrams— Baseball data

Baseball data: (a) alpha vs. (b) correlation ordering



See: `http://www.math.yorku.caSCS/sasmac/corrgram.html`

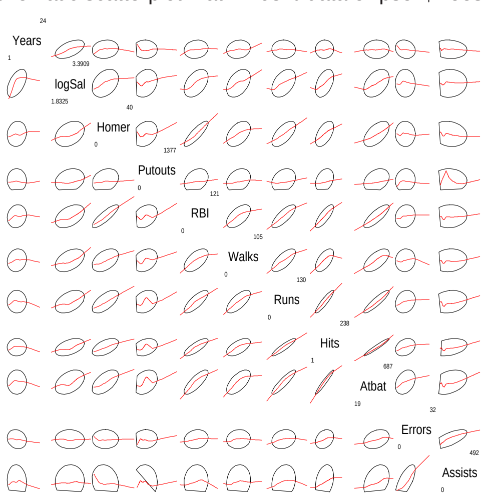
Corrgrams— Auto data



- Correlation ordering shows a coherent pattern
  - Size variables positively correlated
  - Gratio, MPG, repair record positively correlated
  - Negative correlations between the two sets

Corrgrams— Other renderings

Baseball data: schematic scatterplot matrix: 68% data ellipse + loess smooth



- Different renderings for look-up, comparison, detection of patterns, anomalies!

Effect ordering for data displays

- **Information presentation is *always* ordered**—
  - in ***time, or sequence*** (a talk, a written paper),
  - in ***space*** (a table, or graph)
  - Constraints of time and space are dominant— can conceal or reveal the important message.
- **Effect ordering for data display** (Friendly and Kwan, 2003)

Sort the data by the effects to be seen

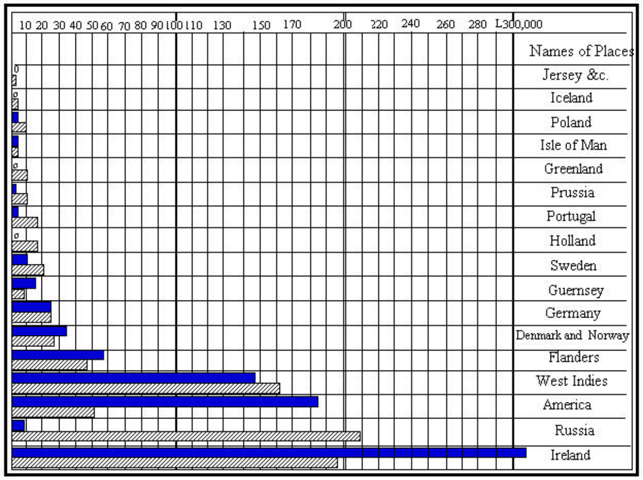
- Applies to:
  - unordered factors for quantitative data
  - categories of variables in frequency tables
  - arrangement of observations and variables in multivariate displays

Effect ordering for data displays

- Multiway quantitative data
  - Main effects ordering— sort unordered factors by means/medians
- Multiway frequency data
  - Association ordering— sort by CA Dim 1 (SVD of residuals from independence)
- Multivariate displays
  - Correlation ordering for variables
  - Clustering/sorting for observations

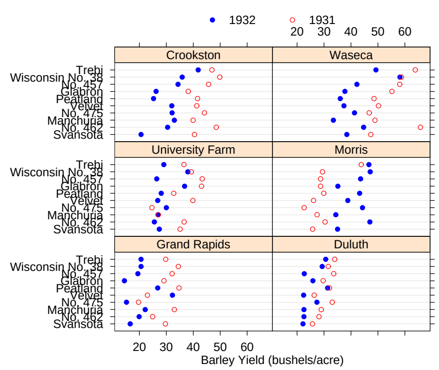
Main effect ordering for tables and charts

Playfair's 1786 barchart of imports and exports of Scotland



Quantitative data: Main effects ordering

- Quantitative response data, cross-classified by one or more factors
- Cleveland (1993)– Barley yields: 10 varieties × 6 sites × 2 years
  - 3-way dot plot, varieties and sites sorted by main effects.
  - All sites except one: higher yields in 1931 than 1932.
  - ⟶ Anomalous site (Morris) might have had years mislabeled.



Main effects ordering: Tabular displays

Average yield (over years) by Variety and Site, ordered ***alphabetically***:

- Good for lookup
- Bad for seeing patterns, trends, anomalies

Table 1: Average Barley Yields (rounded), Means by Site and Variety

Variety	Site: Crookston	Duluth	Grand Rapids	Morris	University Farm	Waseca	***Mean***
Glabron	32	28	22	32	40	46	33.3
Manchuria	36	26	28	31	27	41	31.5
No. 457	40	28	26	36	35	50	35.8
No. 462	40	25	22	39	31	55	35.4
No. 475	38	30	17	33	27	44	31.8
Peatland	33	32	31	37	30	42	34.2
Svansota	31	24	23	30	31	43	30.4
Trebi	44	32	25	45	33	57	39.4
Velvet	37	24	28	32	33	44	33.1
Wisconsin No. 38	43	30	28	38	39	58	39.4
Mean	37.4	28.0	24.9	35.4	32.7	48.1	34.4

Enhanced tabular displays

Average yield (over years) by Variety and Site,

- ordered by ***main effect means***:
- values shaded by (interaction) residual from additive model Yield = Variety + Site
  - Color á la mosaic display: blue for $e_{ij} > 0$, red for $e_{ij} < 0$.
  - Intensity: $|e_{ij}| > \{1, 2\} \times \sqrt{MSE_E}$.

Table 2: Average Barley Yields, sorted by Mean, shaded by residual from the model `Yield = Variety + Site`

Variety	Site: Grand Rapids	Duluth	University Farm	Morris	Crookston	Waseca	***Mean***
Svansota	23	24	31	30	31	43	30.4
Manchuria	28	26	27	31	36	41	31.5
No. 475	17	30	27	33	38	44	31.8
Velvet	28	24	33	32	37	44	33.1
Glabron	22	28	40	32	32	46	33.3
Peatland	31	32	30	37	33	42	34.2
No. 462	22	25	31	39	40	55	35.4
No. 457	26	28	35	36	40	50	35.8
Wisconsin No. 38	28	30	39	38	43	58	39.4
Trebi	25	32	33	45	44	57	39.4
Mean	24.9	28.0	32.7	35.4	37.4	48.1	34.4

Enhanced tabular displays

Yield ***difference*** ($\Delta y_{ij} = 1931 - 1932$) by Variety and Site,

- ordered by ***year effect difference***
- shaded by value ($|\Delta y_{ij}| > \{2, 3\} \times \hat{\sigma}_{\Delta y_{ij}}$)

Table 3: Yield Differences, 1931-1932, sorted by mean difference, and shaded by value

	Site						
Variety	Morris	Duluth	University Farm	Grand Rapids	Waseca	Crookston	***Mean***
No. 475	-22	6	-5	4	6	12	0.1
Wisconsin No. 38	-18	2	1	14	1	14	2.4
Velvet	-13	4	13	-9	13	9	2.9
Peatland	-13	1	5	8	13	16	4.8
Manchuria	-7	6	0	11	15	7	5.5
Trebi	-3	3	7	9	15	5	6.1
Svansota	-9	3	8	13	9	20	7.3
No. 462	-17	6	11	5	21	18	7.4
Glabron	-6	4	6	15	17	12	8.0
No. 457	-15	11	17	13	16	11	8.8
Mean	-12.2	4.6	6.3	8.2	12.5	12.5	5.3

- Negative values for Morris immediately stand out
- Other differences have lower-triangular pattern

Automating main effect ordering: Two-way display

Tukey (1977) two-way display

- Show predicted values and residuals in a two-way table
- Additive model, $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$
- Fitted values, $\widehat{Y}_{ij}$ shown as rectangular grid at coordinates (x, y),

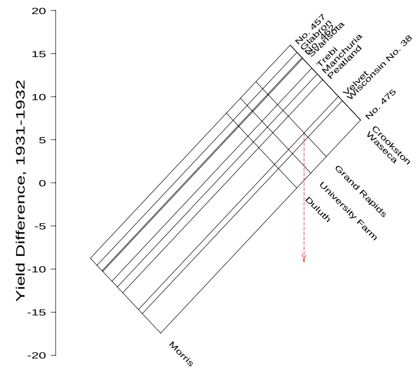
$$x_i = \widehat{\mu} + \widehat{\alpha}_i = \text{row fit}_i$$
$$y_j = \widehat{\beta}_j = \text{col effect}_j$$

- Two-way display ($45°$ rotation) plots:
  - $(x_i + y_j) = \widehat{Y}_{ij}$ = Fit vs.
  - $(x_i - y_j)$ —scaled to keep rectangular
  - $e_{ij} = Y_{ij} - \widehat{Y}_{ij}$ = Residual shown as vectors

Two-way display

Barley yield differences:

- Morris dominates the display
- Residuals, $|e_{ij}| > 2\sqrt{MS_E}$ shown by directed arrows
- Residual for Velvet at Grand Rapids stands out



Effect ordering for frequency tables

Table 4: Hair color - Eye color data: Alpha ordered

	Hair color			
Eye color	Blond	Black	Brown	Red
Blue	94	20	17	84
Brown	7	68	26	119
Green	10	15	14	54
Hazel	16	5	14	29

Table 5: Hair color - Eye color data: Effect ordered

	Hair color			
Eye color	Black	Brown	Red	Blond
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

Model:	*Independence*: [Hair][Eye] χ^2 (9)= 138.29						
Color coding:	<-4	<-2	<-1	0	>1	>2	>4
n in each cell:	$n <$ expected				$n >$ expected		