M O D U L E - I I B

Handling Data

Gather Data- Topical Data- Data Scrapping- Formatting Data, tools- Formatting with code- **Scientific Design Choices using Graphical Display Options - Adding Model with Overlaying (Statistical) Information - Higher-dimensional Displays and Special Structures with parallel coordinates- Scatter Plot Matrices and Mosaic Plot-Time Series and Maps.**

History

Data graphics may be found going very far back in history. but most experts agree Data Visualization really began with the work of Playfair a little more than 200 years ago.

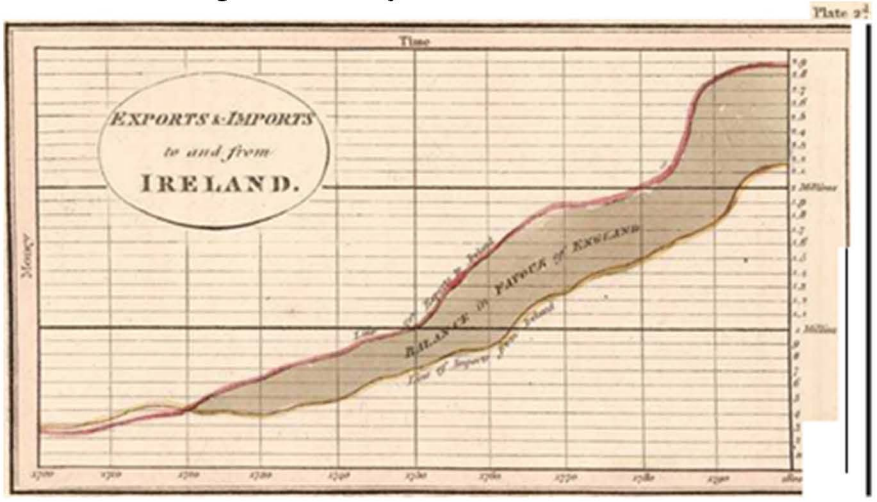
He introduced some modern basic plots (including the barchart and the histogram) and eye-catching displays

Not all his graphics could be described as good, but most were.

A Good Graphic Display

Graphics have been used for a long time to present data

a scanned image from Playfair's Commercial and Political Atlas of 1801



Presentation describes the fairly continuous increase of both imports and exports, and the fact that the balance was in favour of England from 1720 onward,

The context can be seen easily.

Some improvements might be made, but overall, the display is effective and well-drawn.

What makes graphic display a one good?

- ☞ In any successful graphic there must be an effective blending of content, context, construction and design.
- ☞ Handling the data in such a way that it becomes easier for people to understand and comprehend the given information.

Content, Context and Construction

Content, Context and Construction

What (information) is plotted comes first (important),

without content no amount of clever design can bring meaning to a display.

A good graphic will convey information

but a graphic is always part of a larger whole

the context, which provides relevance to the graphic.

So,

A good graphic has to complement other related material and fit in

in terms of

content

style &

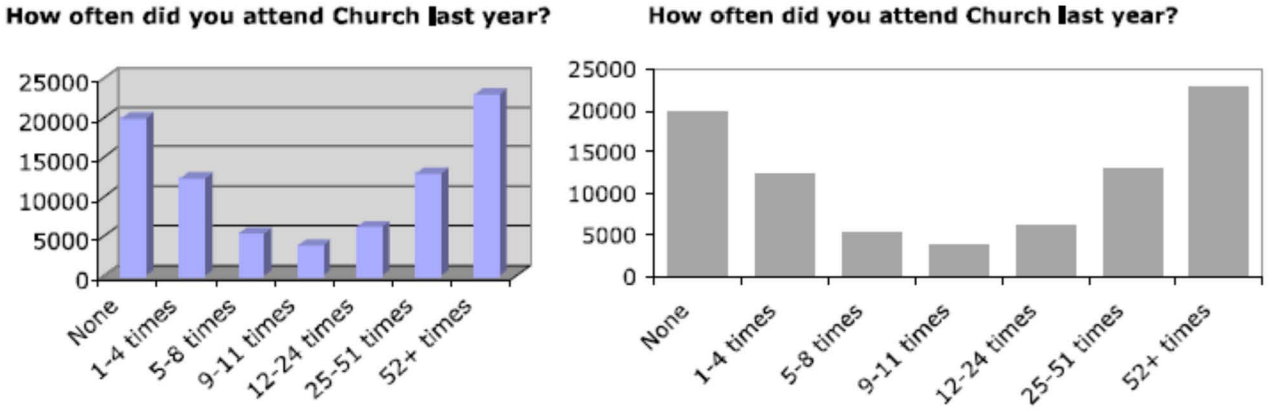
layout.

Finally,

a graphic constructed and drawn well, will look good.

Example

two similar displays of the same data - how often they had attended church in the last year.



Left-hand graph

Includes gridlines and a coloured background and uses 3-d columns to represent the data counts

Right-hand graph

Sticks to basics.

In general,

The right-hand display is to be preferred

3-d columns can cross gridlines

Zero values would be misleadingly displayed.

Both graphics convey the same overall information.

The potential weakness in both graphics is the set of categories.

Grouping the data together in different ways could give quite different impressions.

It is advisable to think hard about what should be shown and to check with others if the graphic makes the desired impression

Presentation Graphics and Exploratory Graphics

Two main reasons for using graphic displays of datasets

- To present or
- To explore data

Presenting data

Involves

Deciding what information to convey
Drawing an appropriate display for the content
Appropriate presentation for the intended audience.

Also think about

How the plot might be perceived &
Whether it will be correctly understood

So great care should be taken in preparing the most appropriate display.

Exploring data

More individual in nature,
Using graphics to find information and generate ideas.
Involves

Drawing many displays
Changing discarding redrawing drawings
Each drawing having a short life span

Principles & guidelines for good presentation graphics

Have a very small role

Personal taste & individual working style

Play important role

Presentation
(What to Whom, How and Why)

Is it possible to make a mess of presenting simple statistical information?
Yes,
Technically there is much that can go wrong.

distortions
misleading scales
3-D displays of 2-D data
Difficult to make fair comparisons
areas un-proportional to values
too much information is crammed into a small space

semantically

The caption, the headline & the accompanying article can tell different stories (ideally all three should have same context)

Whether or not a graphic is successful as a display depends on its
- subject,
- context &
- aesthetic considerations

=>It depends on
- what it is supposed to show,
- what form is chosen &
- its audience.

Scientific Design Choices

Plotting a **single variable** should be fairly easy.

The type of variable will influence the type of graphic chosen.

Example,
- Continuous variables : histograms or boxplots
- Categorical variables : barcharts or piecharts

Whether the data should be
- Transformed or
- Aggregated

Will depend on the
- Distribution of the data &
- Goal of the graphic.

Scaling and captioning is straightforward, and to be chosen with care.

Multivariate graphics is not simple

The main decision is
- The form of display,
- The choice of variables &
- Variable's ordering

Example
- Dependent variable should be plotted last
- In a scatterplot it is traditional to plot the dependent variable on the vertical axis.

Choice of Graphical Form

The choice display depends on the type of data to be displayed

There are many display options

- Bar charts,
- Pie charts,
- Histograms,
- Dot plots,
- Boxplots,
- Scatterplots,
- Rose plots,
- Mosaic plots

And many other kinds of data display

Example:

Pie charts are good for displaying shares for a small number of categories

Boxplots are good for emphasizing outliers

While

Univariate continuous data cannot be displayed in a pie chart

Bivariate categorical data cannot be displayed in a boxplot

A poor choice graph type cannot be rectified layer by other means so it is important to get it right at the start.

There are always many multiple alternative options to consider which can be equally good or good in other ways.

Example: emphasizing different aspects of the same data.

Simply adopting the default of whatever computer software is suggesting unlikely to be wise.

Graphical Display Options

Scales

Defining the scale for the axis for

A categorical variable may depend on what the categories represent or on their relative sizes.

A continuous variable it is more difficult and dependent on the endpoints, divisions and tick marks chosen.

Nice' scales should possess the properties

- Simplicity
- Granularity
- Coverage, &
- Zero (if possible)

The user must also check the scale and be prepared to amend it according to the data

Simple answer; to choose scales running from the minimum to the maximum

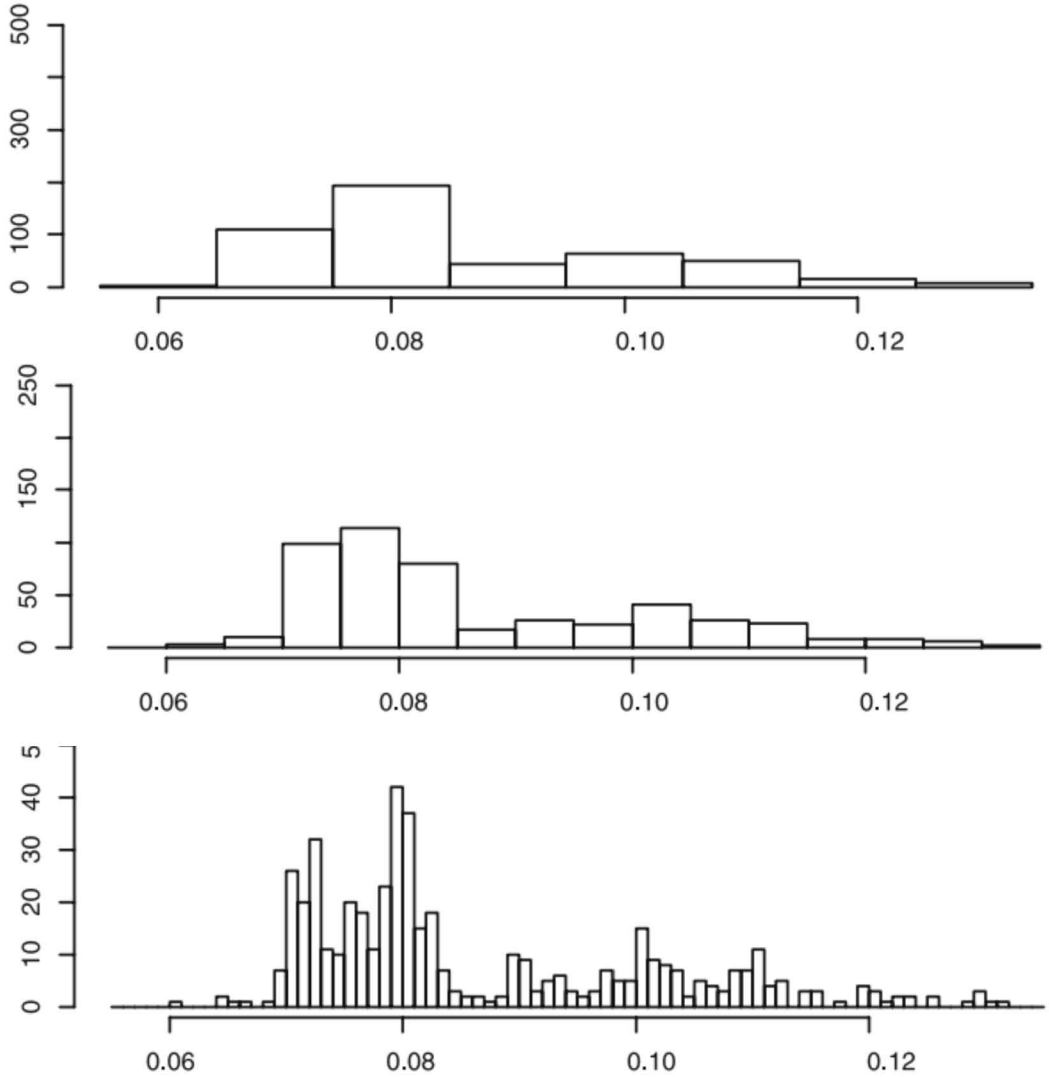
- ⇨ Points on the boundaries and may be obscured (not properly visible/observed)
  - ⇨ it is good practice to extend the scales beyond the observed limits to ensure readability and understandability.

There is no compulsion to include zero in the scale, but there should be a reason for not including it.

Most common reasons

Lowest value can be chosen as the start of the scale instead of 0

Example: Hidalgo stamp thickness data different bandwidths provide different views.



Note : The third histogram was used for publication

Display Issues to note

Pros

All three plots be aligned exactly and have the same total area Common scaling is used in one form or another Horizontal value axis

cons

would be nicer if it extended further to the right.

Sorting and Ordering

Display can be influenced by many factors.

When plotting more than one variable

- The position
- The order

In the graphic can makes huge differences.

Ordering

Choice of ordering can dependent on the goal of the graphic

Good ordering can help understand graphic better, portray the Purposes

Some standard ordering

alphabetic ordering

a standard default

appropriate for comparison

geographic or other grouping

Eg.. Shares by market sector

ordered by size

ordered by a secondary variable

Example: two bar charts of the same data,

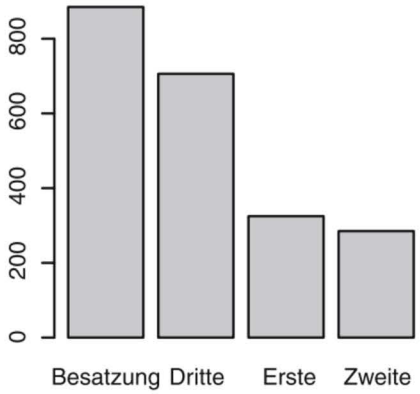
the number of passengers in each class and the crew on the titanic

Besatzung => Crew (908),

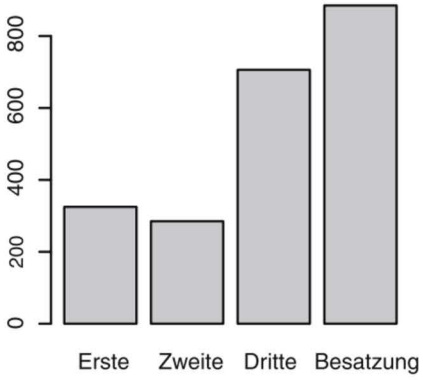
Erste =>First Class (325),

Zweite =>Second Class(285),

Dritte =>Third Class(706),



Ordered alphabetically (German)



Ordered by class

Adding Model/ (Statistical) Information

Guides may be drawn on a plot

As a form of annotation

Useful for

Emphasising particular issues

Say which values are positive or negative

Sloping guides

Highlight deviations from linearity

Fitted lines

Eg..

Polynomial regression or

Smoothers,

May be superimposed on data

To show

The hypothesised overall structure

Highlight local variability and

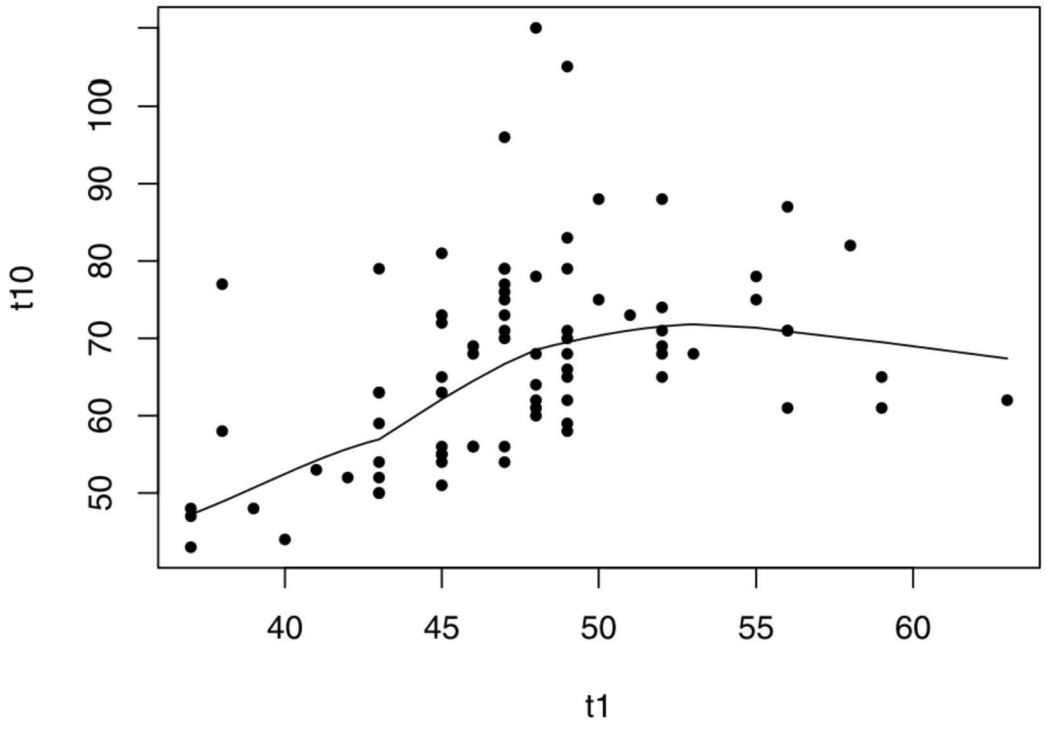
Any lack of fit.

Example:

Times from the first and last stages of a 100km road race.

A lowess smoother has been drawn.

Annotation suggests that there is a linear relationship for the faster runners and a flat one for the slower ones.



Captions, Legends and Annotations

Captions

Caption should fully explain the graphic and the source of the data

Since

Very long captions are probably not liked by the readers

Caption may not fulfil its role

A preferable solution can be a compromise where

The caption outlines the information in the graphic &

A more detailed description can be provided in the text

Note;

Graphics which require extensive commentary may be trying to present too much information at one go.

Legends

Legends provide description about symbols and/or colours used in the graphics

It shall be preferred if the information should be on the plot directly and not in a separate legend

Annotations

Annotations are used to highlight particular features of a graphic

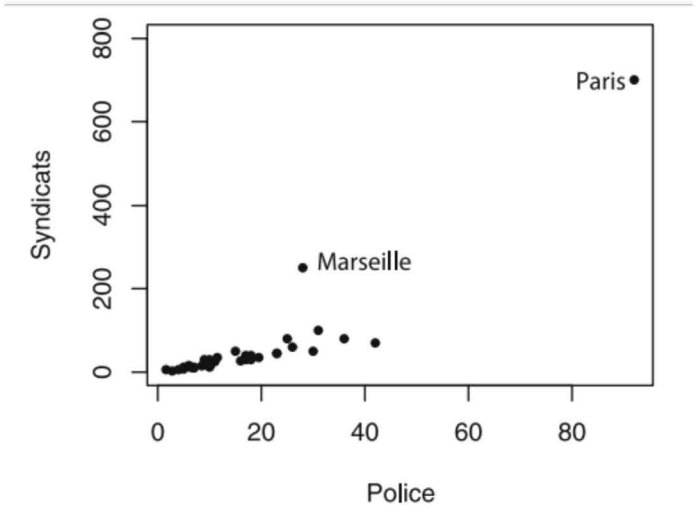
useful for identifying events in time series

drawing attention to particular points in scatterplots.

Example:

Scatter plot describing union estimate and police estimate of number of protesters

Annotations drawing attention to Paris & Marseille where the difference is large



Positioning in Text

Graphics and related text must be on the same page or on facing pages
As it is inconvenient to have to turn pages back and forth if graphics and its related text are on different pages

However, it is not always possible to avoid this.

Size, Frames and Aspect Ratio

Size

Graphics should be large enough for the reader to see the information in them clearly and not much larger.
Size of the graphic should be relative to the surrounding layout.

Frames

Frames may be drawn to surround graphics

Frames take up space and add to the clutter

Hence,

Frames should be used only for purposes of separation,

I.e.

Separating the graphic from other graphics or
Separating the graphic from the text.

Aspect ratios

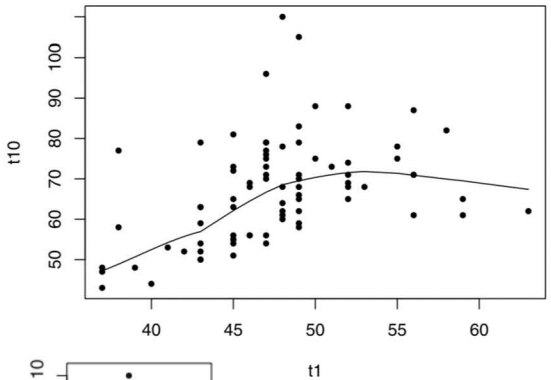
Aspect ratios have a strong effect on the perception of graphics.
For time series to show

Gradual change

Grow the horizontal axis and shrink the vertical axis.

Dramatic change

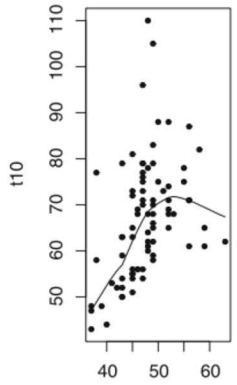
Grow the vertical axis and shrink the horizontal axis.

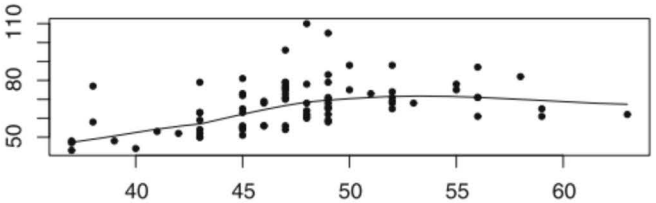


Times from the first and last stages of a 100km road

y (vertical) axis shrunk to project a gradual change ⬇

⬅X (horizontal) axis shrunk to project a dramatic change





Colour

Colour is potentially one of the most effective ways of displaying data.
In practice it is also one of the most difficult to get right.

The chosen colour schemes should

Blend well & distinguish

Between different categories.

Have to bear in mind:

Some people are colour blind;
Colours have particular associations (red for danger or for losses);
Some colours may not be reproduced in print
Colour can be a matter of personal taste.

Higher-dimensional Displays and Special Structures

Scatterplot Matrices (Sploms)

Plotting each continuous variable against every other variable is effective for small numbers, but can be a complex task for larger number of variables.

This Example:

Displays the data from emissions tests of 381 cars sold in Germany.

It plots

Engine size,
Performance
Fuel consumption
Pollutants &
CO2

It reveals

Engine size, performance and fuel consumption are

Linearly related

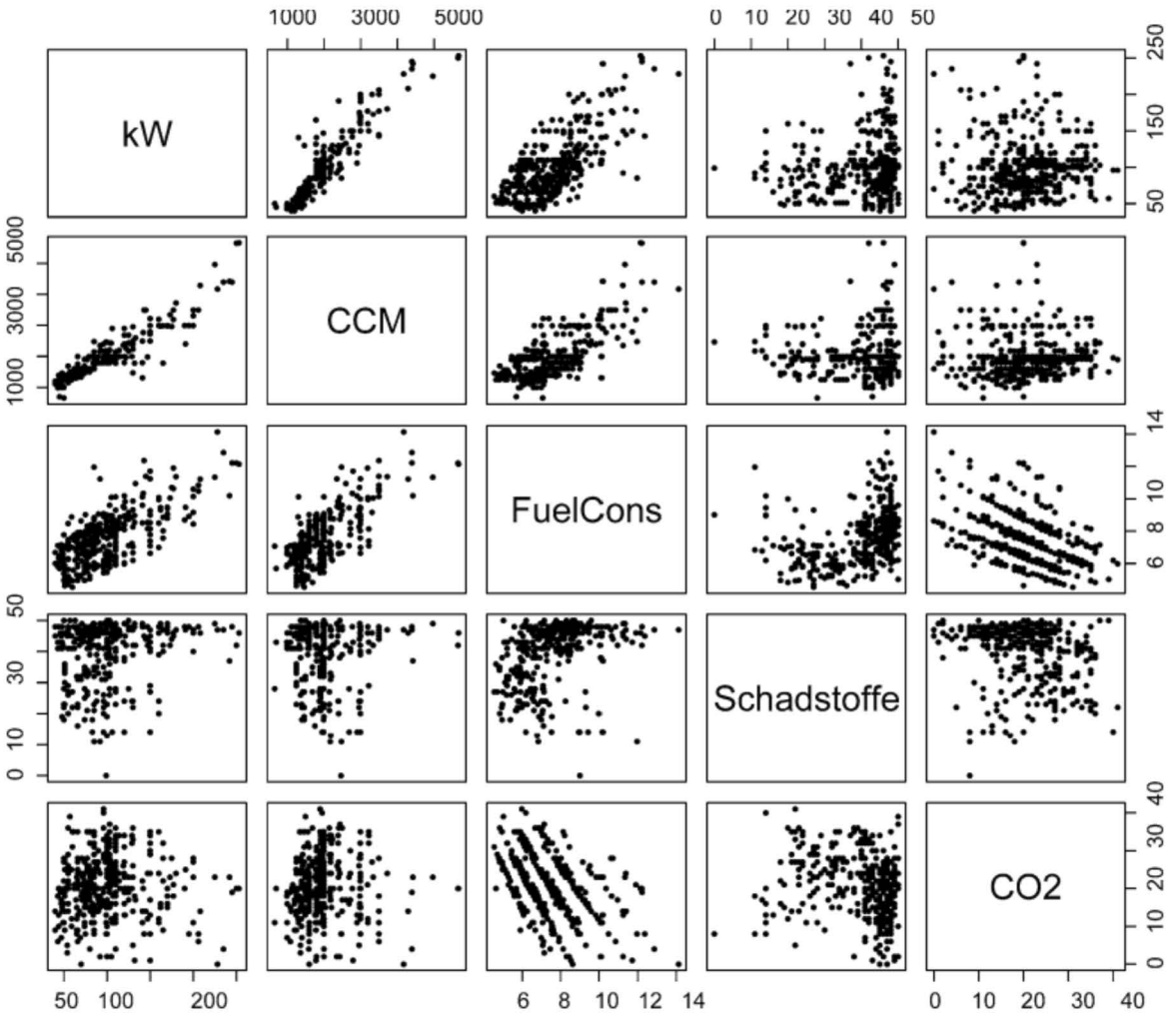
While

Pollutants CO2 and fuel consumption are

Negatively correlated

While packing so many plots into a small space

It is important to cut down on scales.
Variable names can be placed on X / Y or on the diagonal
(this example placed variable names on the diagonal)
The same can be plotted with histograms of the individual variables



Parallel Coordinates

Parallel coordinate plots are valuable for simultaneously displaying large numbers of continuous variables.

Showing too much information at once has several implications:

- Not all information will be visible in one plot
  - (so that several may be needed)
- Formatting and scaling will have a big influence on
  - What can be seen &
  - What remains unseen
- Some overlapping is inevitable
  - Need good blending methods
  - Good density estimation methods.

Example

Plots the cumulative times of the 147 cyclists at the ends of the 21 stages of the 2004 Tour de France.

- The axes all have the same scale
  - So that differences are comparable.
- The best riders take the shortest time and are at the bottom of the plot.

The axes have been aligned at their means,

Without some alignment graphic would be unclear.

Blending has been applied to reduce the overprinting in the early sprint stages where all riders had almost the same times.

If more blending is used then the individual lines for the riders in the later stages of the race become too faint.

This single display conveys a great deal about the race.

In the early stages at most a few minutes separates the riders.

On the mountain stages there are much larger differences and individual riders gain both time and places

(where a line segment crosses many others downwards)

Note that there are relatively few line crossings over the later stages of the race,

Which means, perhaps surprisingly, that not many riders changed their race ranking.

This graphic might be improved in a number of ways:

The axes could be labelled

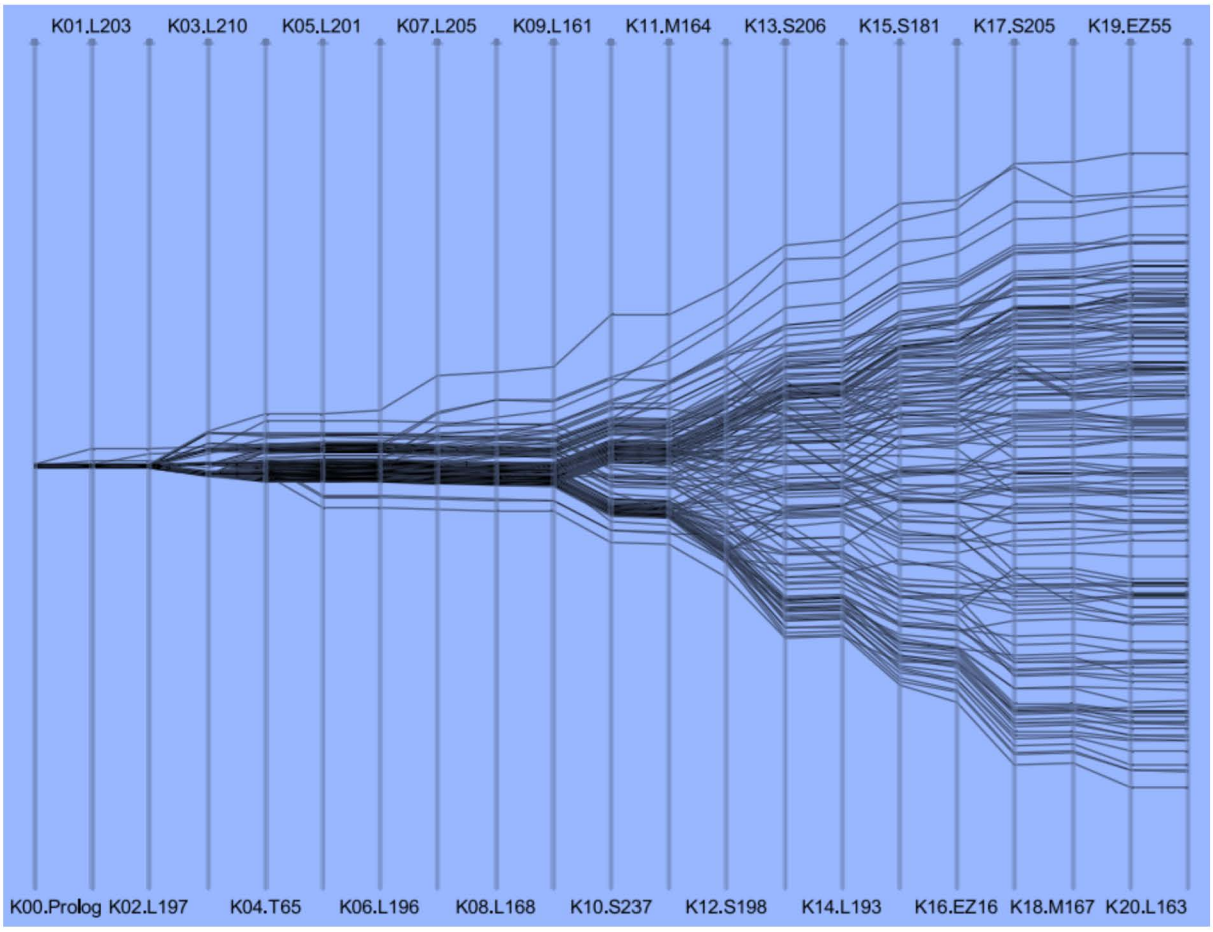
(though there is little space for this)

The vertical axes could be drawn less strongly

Scale information could be added

(the range of the vertical axes is about four hours)

The level of -blending might be varied across the display.



Mosaic Plots

Mosaic plots display the counts in multivariate contingency tables.

Example

The five-dimensional mosaic plot displays the data are from patterns of arrest based on 5226 cases in Toronto.

Each column represents one combination of the four binary variables Gender, Employed, Citizen, and Colour.

The width of a column is proportional to the number with that combination of factors.

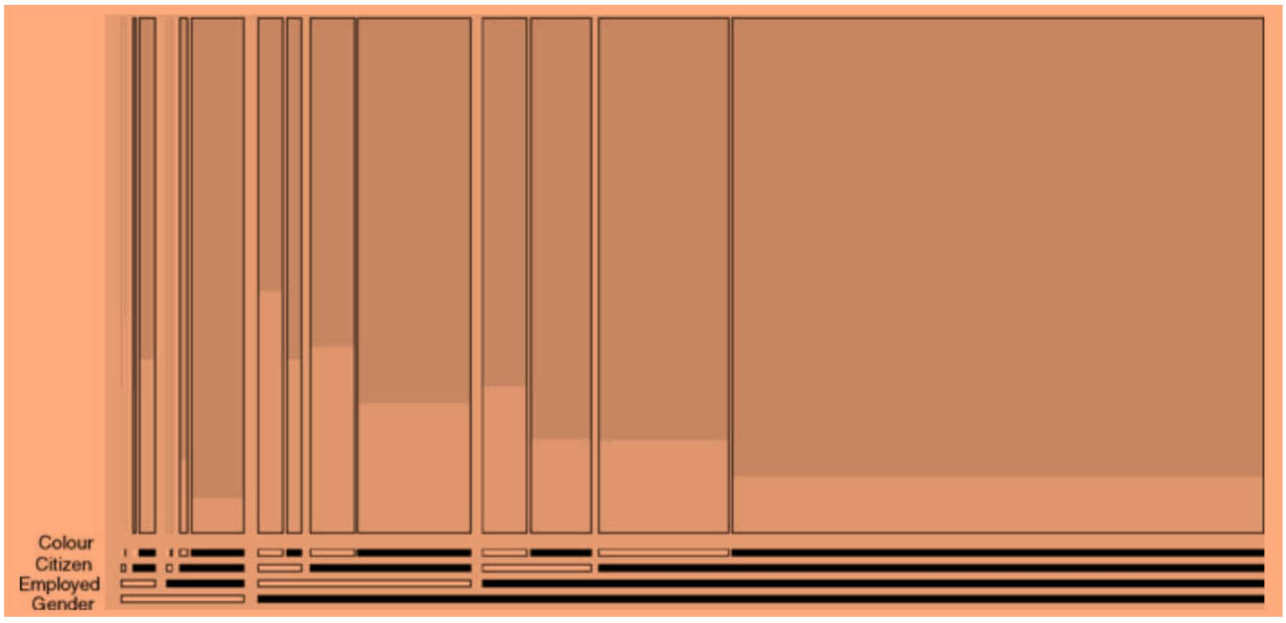
Those stopped who were not released later have been highlighted.

- Over 90% of those stopped were male.
- Some of the numbers of females in the possible eight combinations are too small to draw firm conclusions.
- Each pair of columns represents colour and the proportion not released
- Amongst the males lower amongst the whites for all combinations of other factors.
- The general decline in the level of highlighting across the male columns shows that the proportion not released is lower if the person is a citizen and lower still if they are employed

Example shows the difficulties in displaying data of this kind in a graphic for presentation

Colour, aspect ratio and size can make a big difference

Labelling is the main problem.



Small Multiples Displays

A set of smaller plots can be used to avoid overloading a single large plot with too much information

Set of smaller & comparable plots can be effective for subgroup analyses

Example

The graphic depicts the fuel consumption of cars

The boxplots show that diesel cars have generally lower fuel consumption

The bar chart on the left shows that very few of (natural) gas and hybrid groups were measured

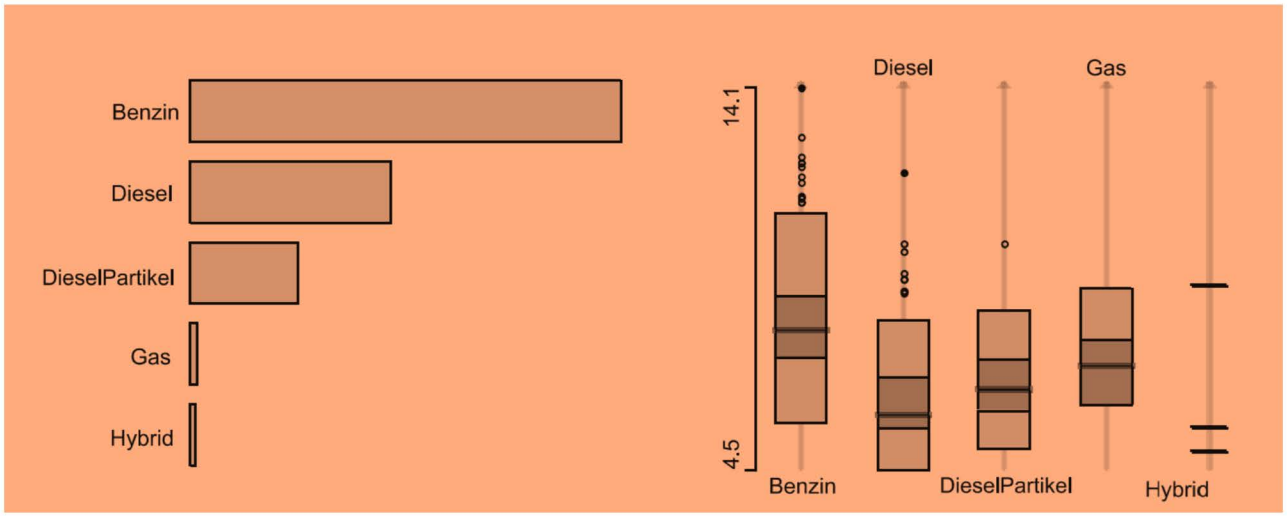
Small groups are always a problem.

It should a be noted that

Other cars were measured by litres/100km, while natural gas cars were measured by kg/100km

Careful captioning is necessary to ensure which smaller plot is which

Common scaling is essential.



Trellis Displays

Example

is a trellis display of emissions' data for the 374 petrol or diesel cars.

cars have been grouped by

engine type (the rows) &

engine size (the columns).

An equal count grouping has been used for engine size,

=> the shaded parts of the cc bars have different lengths

Observations

Engine size seems to make little difference

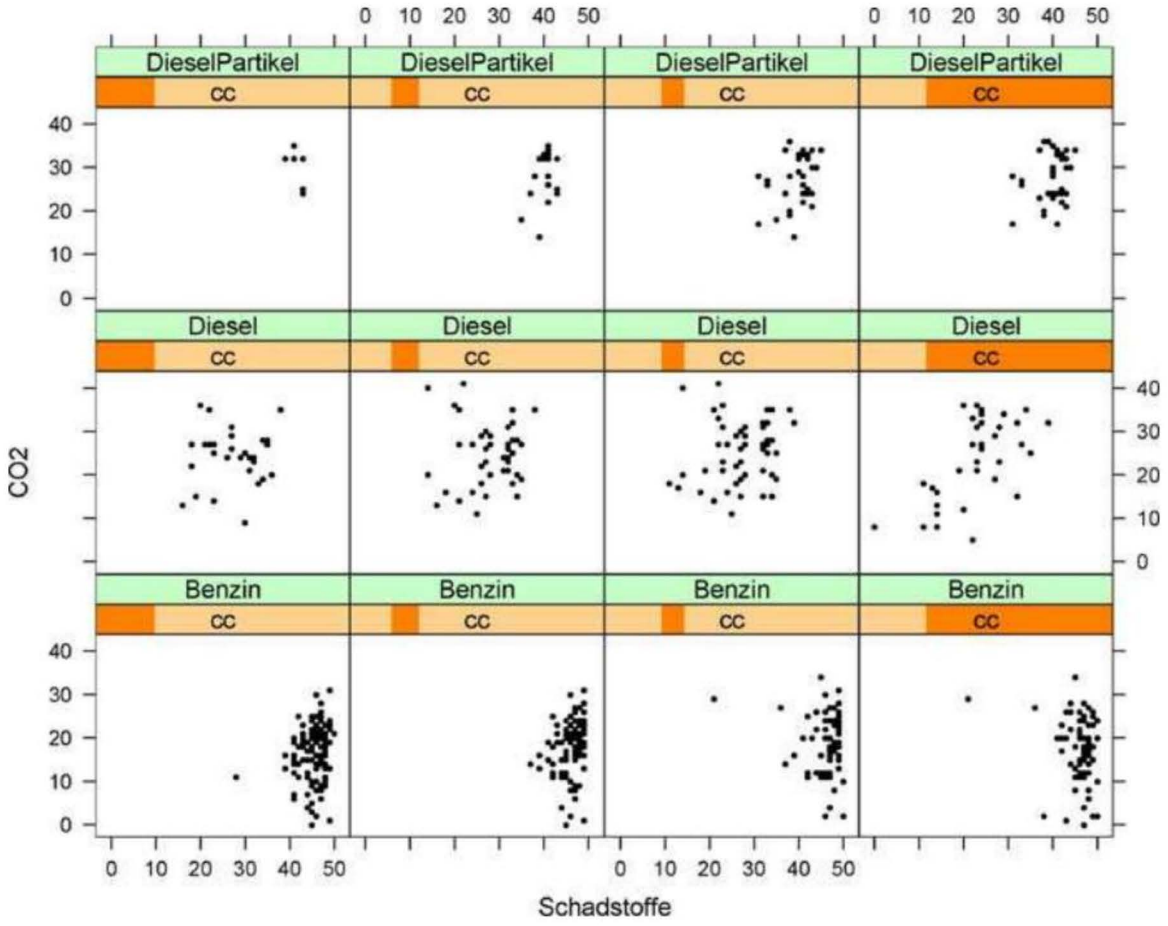
as the plots in each row are similar to one another.

The type of engine makes more difference

with diesel engines showing significant difference from the other two types.

There are a few local outliers amongst the petrol cars.

When several plots of the same kind are displayed,
they can be plots of subsets of the same data or
plots of different variables for the same dataset



Time Series and Maps

Time Series

Time series are special because of the strict ordering of the data, and good displays respect temporal ordering.

It is useful to differentiate between value measurements at particular time points and summary measurements over a period

e.g.

a patient's weight or a share price (time point measurements)

how much patient ate in the last month (summary measurements)

Time scales have to be carefully chosen.

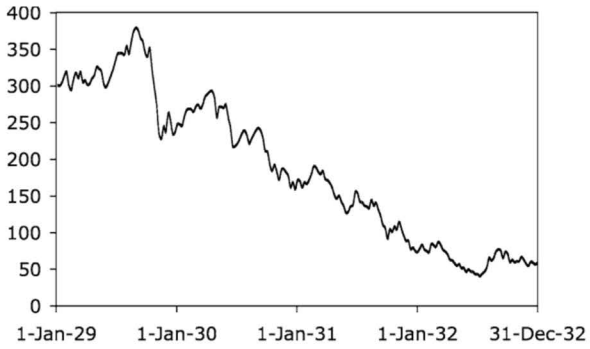
The choice of time origin (start time)is particularly important

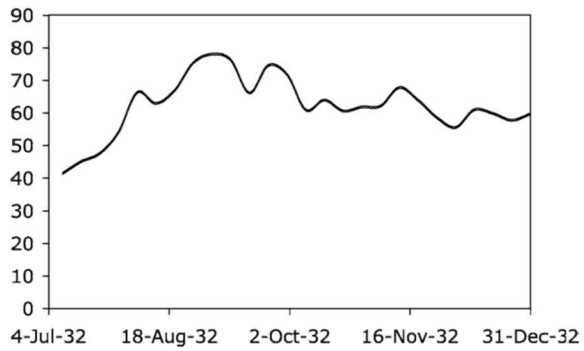
Time points for value measurements may not match the calendar scale

Time units for summary measurements may be of unequal length

The time period chosen and the aspect ratio used for a time series plot can make a big difference in the interpretation of the data

If several time series are plotted in the same display, then it is necessary to ensure that they are properly aligned in time





Maps

graphical displays can be very informative, but Geographic data are complex to analyse

The main problems are

areas do not reflect the relative importance of regions

spatial distance is not directly associated with similarity or nearness

need clarity on

how to use colour scales to show values &

how to choose colour schemes

Example

shows that cancer rates are highest along the East Coast and lowest in the Midwest

State Economic Areas (SEAs) have been chosen because using states smooths the data

(consider Nevada in the West with its high cancer rate around Las Vegas, but its lower rate elsewhere)

The map on the website is in colour,

on a scale from deep red for high rates to dark blue for low.

this would not reproduce well in a grey-scale view

so the alternative print version that is used here