MAS465 Multivariate Data Analysis
MAS6011(1) Dependent Data

2. Data visualisation

School of Mathematics and Statistics

Dr Frazer Jarvis (J12)
a.f.jarvis@sheffield.ac.uk

Plan for these slides

- Data visualisation
- Graphics in R
  - 1 dimension
  - 2 dimensions
  - several dimensions

Data visualisation

One could give a whole course on data visualisation (or datavis as it is abbreviated), and many books have been written on the subject. We will barely scratch the surface, focusing on the capabilities of R!

When you receive a data set to analyse, your first task is to try to spot the patterns and other interesting features in the set.

A useful R command is `summary()` which gives produces several summary statistics for the data set (although, curiously, not the variance matrix).

In these slides, we'll use the data set `airpoll`. Instructions for downloading and opening it in an R session are given in the notes.

Data visualisation

`airpoll`

The data set consists of seven variables (`Rainfall`, `Education`, `Popden`, `Nonwhite`, `NOX`, `SO2`, `Mortality`) for each of 60 US cities, and the hope is to understand the relationship between mortality rates (the variable `Mortality`) and the other variables, especially the amount of sulphur dioxide in the air (`SO2`).

The `summary` command produces columns giving the minimum, maximum, quartiles, median and mean values of each of the seven variables.

Data visualisation

But to see interesting patterns, and to spot e.g., outliers, it would be much better to get some pictorial representation of the data.

Beware that humans are very good – too good! – at spotting patterns in pictorial data, and may see patterns where none exist; this is why a statistical analysis, if available, should also be carried out.

Data visualisation

In 1 dimension, you are all familiar with the options for drawing plots:

- With a fairly small number of points:
  - 1-dimensional scatter plots (i.e., dot plots)
- With a larger number of points:
  - stem-and-leaf plots
  - histograms
  - box plots

Care needs to be taken over the latter plots, so that interesting information (e.g., bimodality, outliers) is not missed.

Data visualisation

In 2 dimensions, there are similar tools one can use.

- With a fairly small number of points (e.g., `airpoll`):
  - scatter plots
- With a larger number of points:
  - bivariate histograms
  - bivariate box plots

In a bivariate histogram, the frequency of each bin has to be plotted vertically, making a 3-dimensional shape. For a bivariate boxplot (see also a following slide), we plot ellipses centred on the mean, containing suitable proportions of the data.

Graphics in R

We illustrate some of the possible enhancements to the basic scatterplot in the next slides.

Given the data set `airpoll`, type

```
attach(airpoll)
```

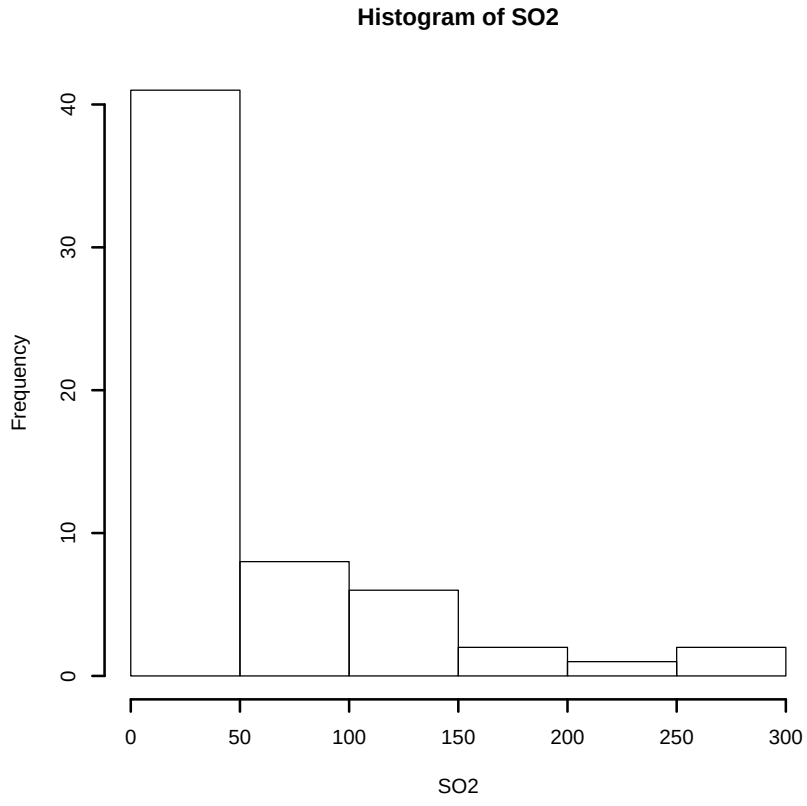
to read it into R's path.

We'll remind you of the commands for 1-dimensional data (histograms and boxplots), before moving onto 2- and higher-dimensional data.

Graphics in R – 1 dimension

Histogram of SO2

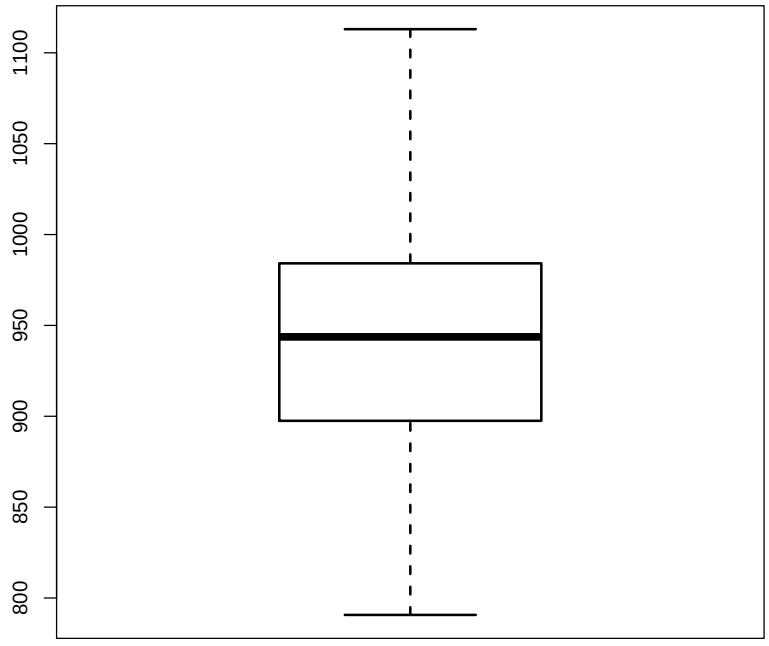
```
hist(SO2,lwd=2)
```



Graphics in R – 1 dimension

Boxplot of `Mortality`

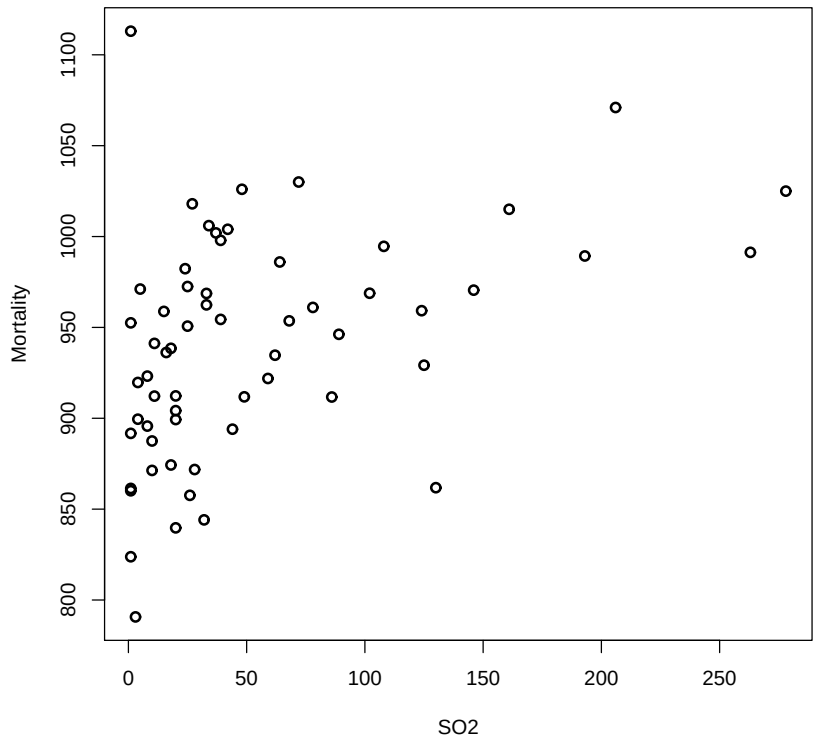
```
boxplot(Mortality,lwd=2)
```



Graphics in R – 2 dimensions

Basic plot. . .

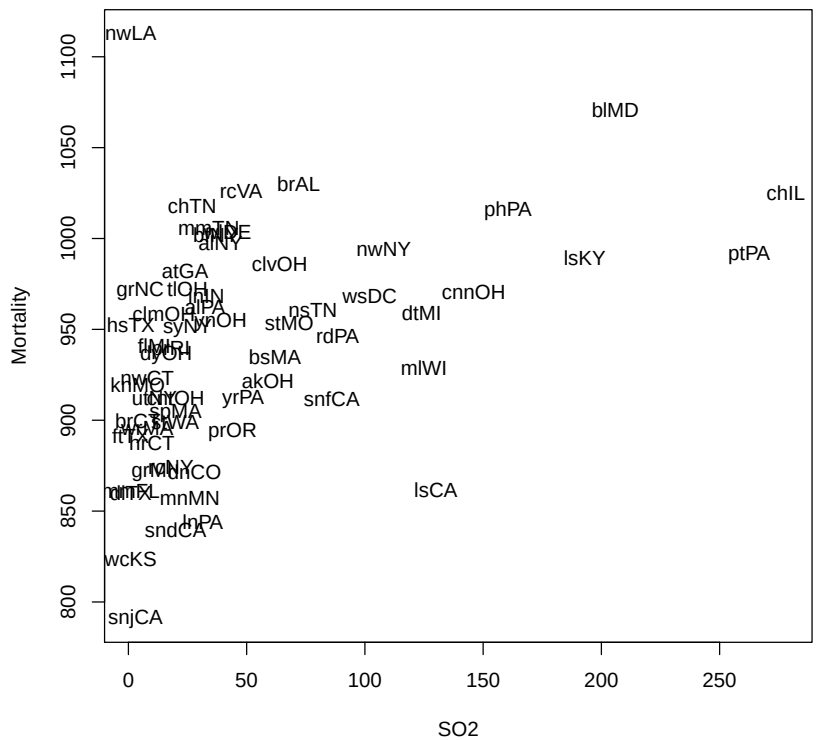
```
plot(SO2,Mortality,pch=1,lwd=2)
```



Graphics in R – 2 dimensions

with names

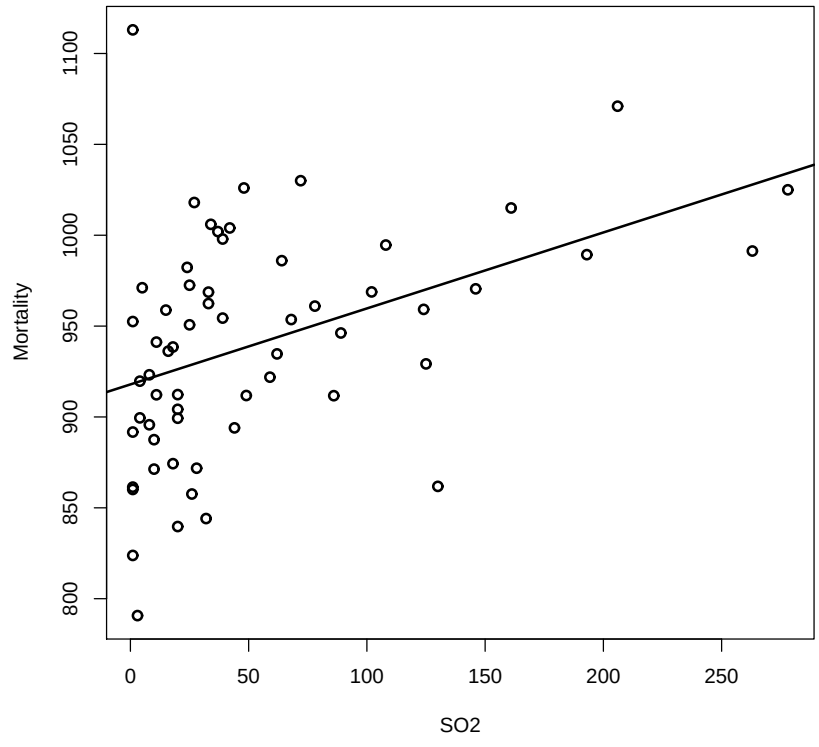
```
names<-abbreviate(row.names(airpoll))
plot(SO2,Mortality,lwd=2,type="n")
text(SO2,Mortality,labels=names,lwd=2)
```



Graphics in R – 2 dimensions

or a regression line. . .

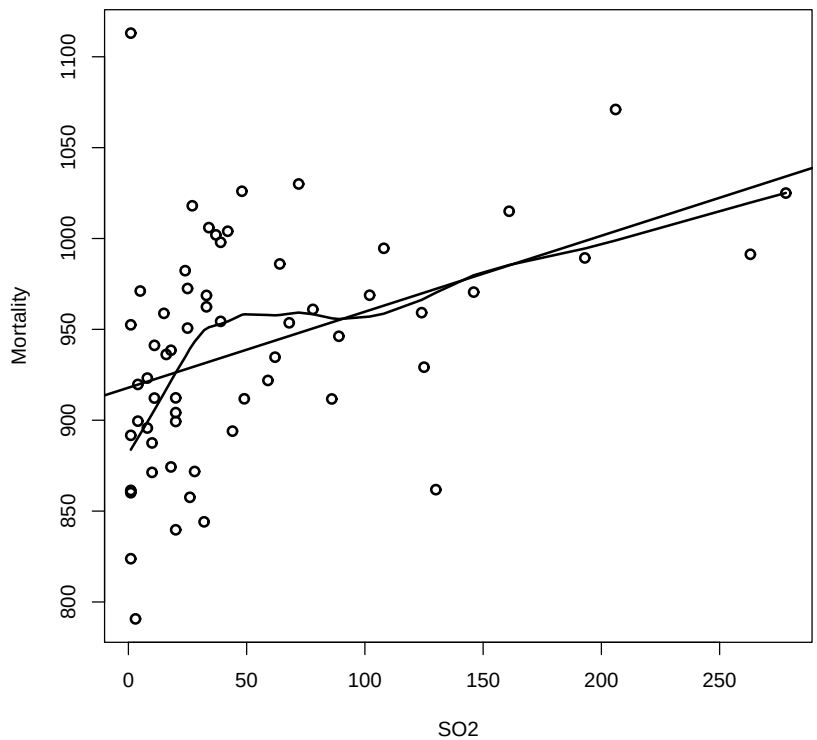
```
plot(SO2,Mortality,pch=1,lwd=2)
abline(lm(Mortality~SO2),lwd=2)
```



Graphics in R – 2 dimensions

with a local best fit

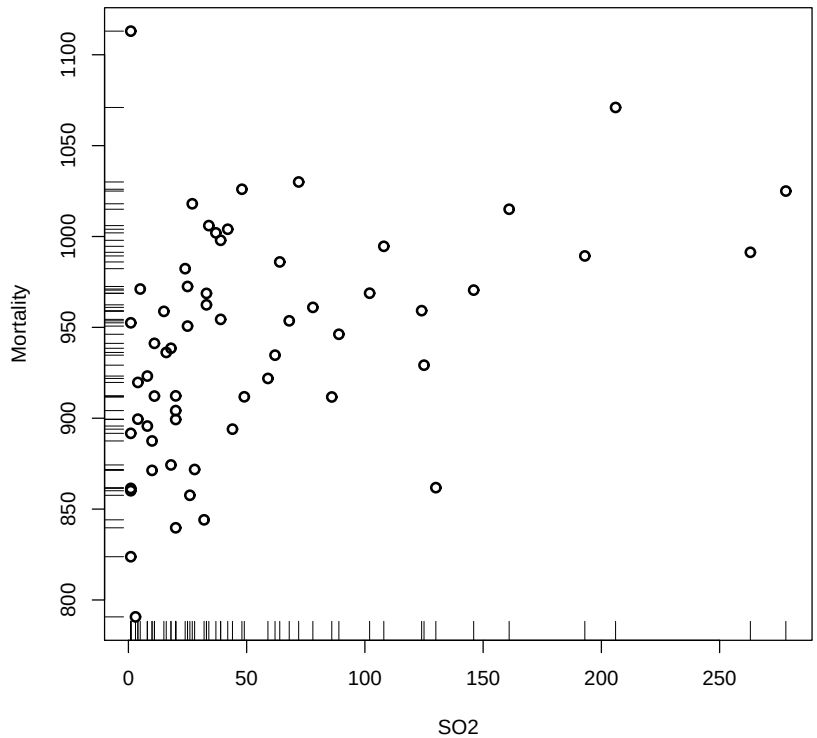
```
plot(SO2,Mortality,pch=1,lwd=2)
abline(lm(Mortality~SO2),lwd=2)
lines(lowess(SO2,Mortality),lwd=2)
```



Graphics in R – 2 dimensions

with marginal distributions

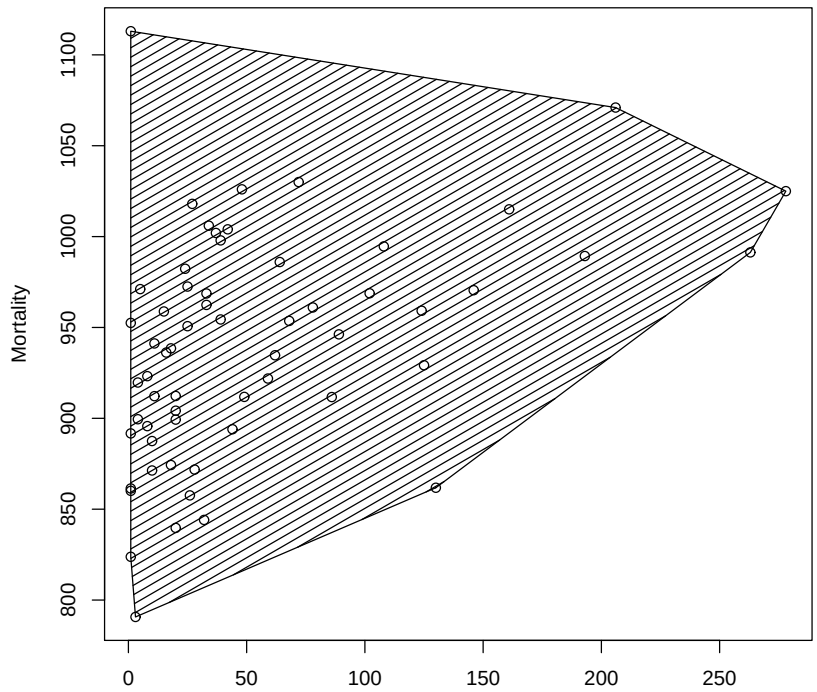
```
plot(SO2,Mortality,pch=1,lwd=2)
rug(SO2,side=1)
rug(Mortality,side=2)
```



Graphics in R – 2 dimensions

or a convex hull

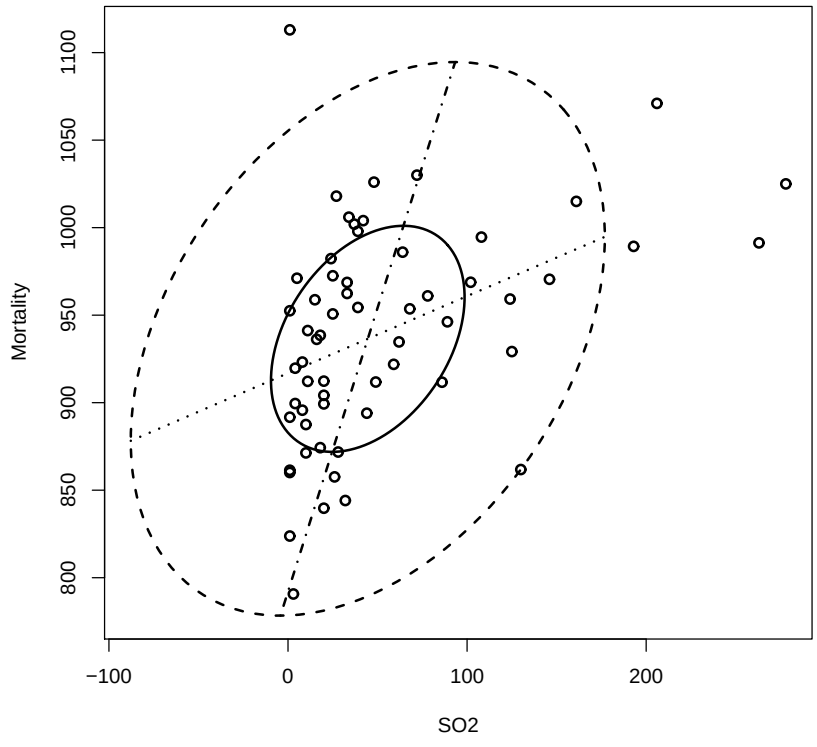
```
hull<-chull(SO2,Mortality)
plot(SO2,Mortality,pch=1,lwd=2)
polygon(SO2[hull],Mortality[hull],density=15,
  angle=30)
```



Graphics in R – 2 dimensions

Everitt's bivariate boxplot bvbox

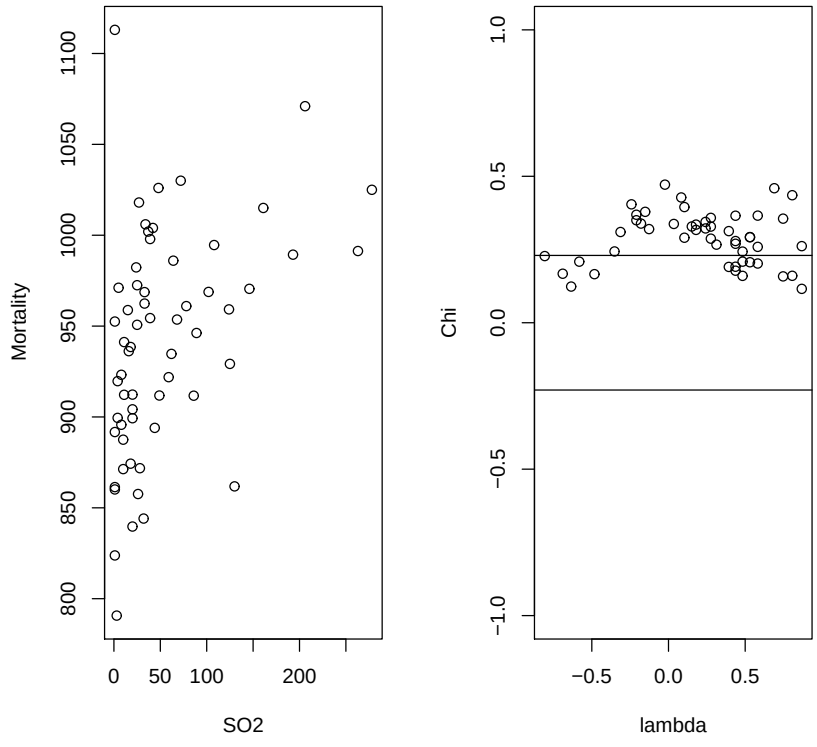
```
bvbox(cbind(SO2,Mortality),xlab="SO2",
  ylab="Mortality")
```



Graphics in R – 2 dimensions

Everitt's `chiplot`

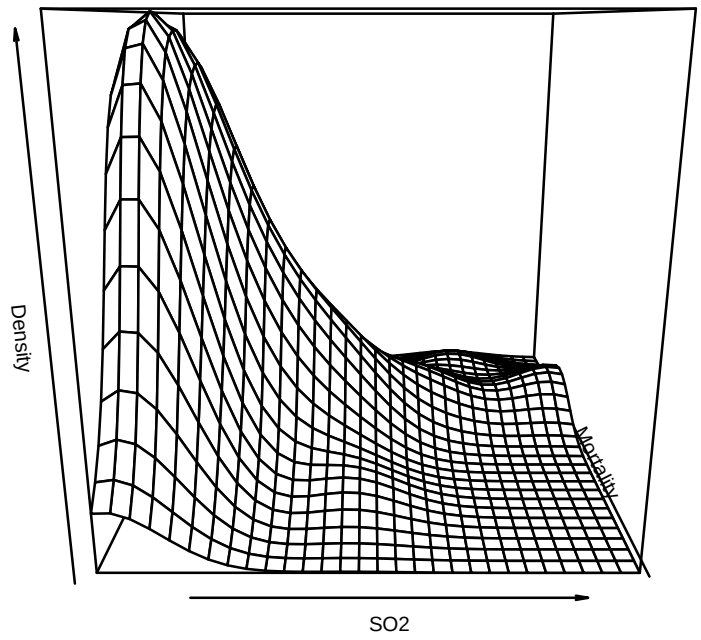
```
chiplot(SO2,Mortality,
  vlabs=c("SO2","Mortality"))
```



Graphics in R – 2 dimensions

Three-dimensional views: a density plot

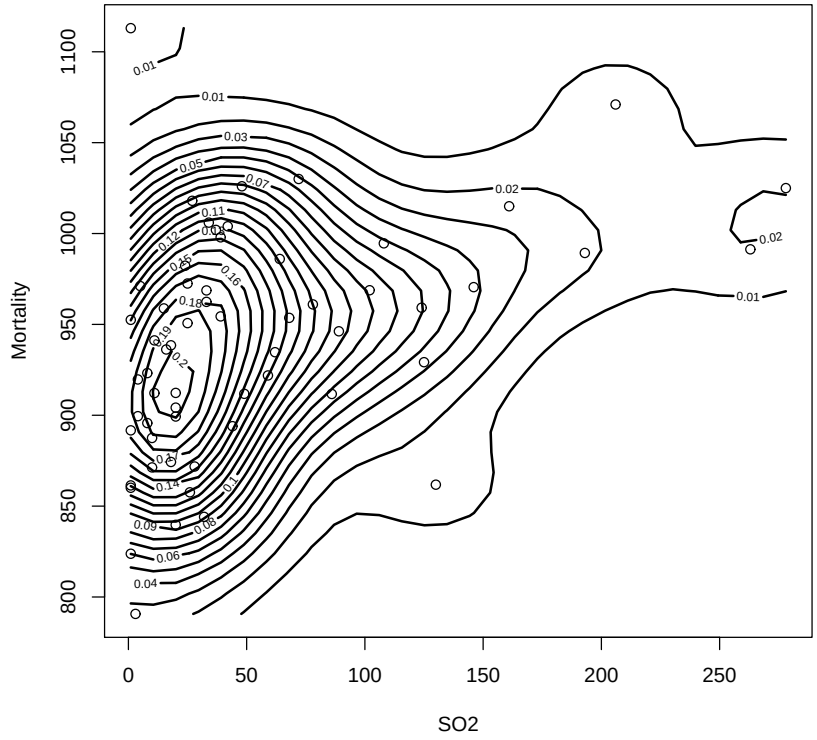
(code in notes)



Graphics in R – 2 dimensions

Three-dimensional views: a contour plot

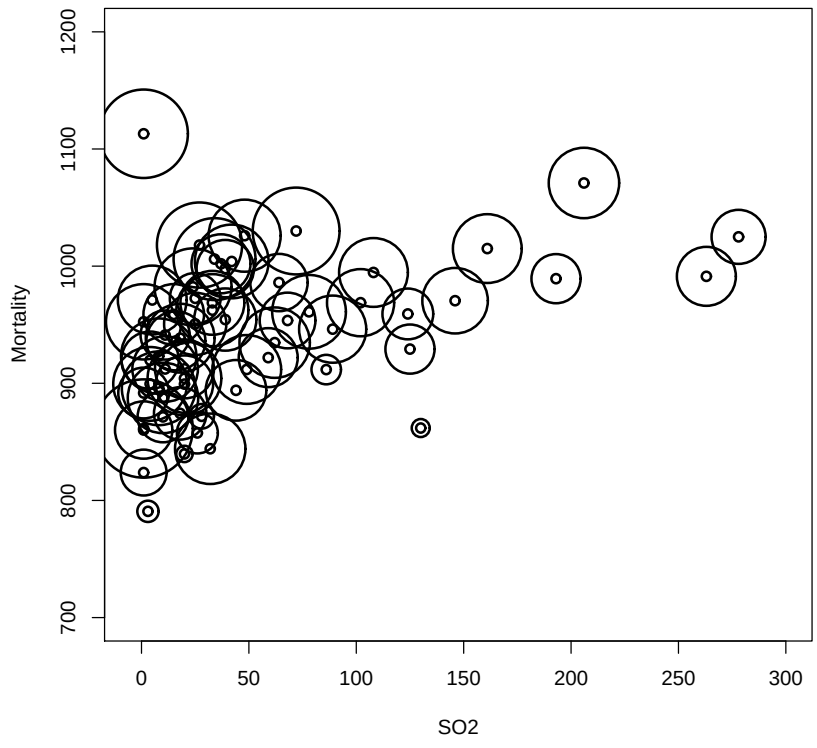
(code in notes)



Graphics in R – 3 dimensions

A third variable – the bubbleplot

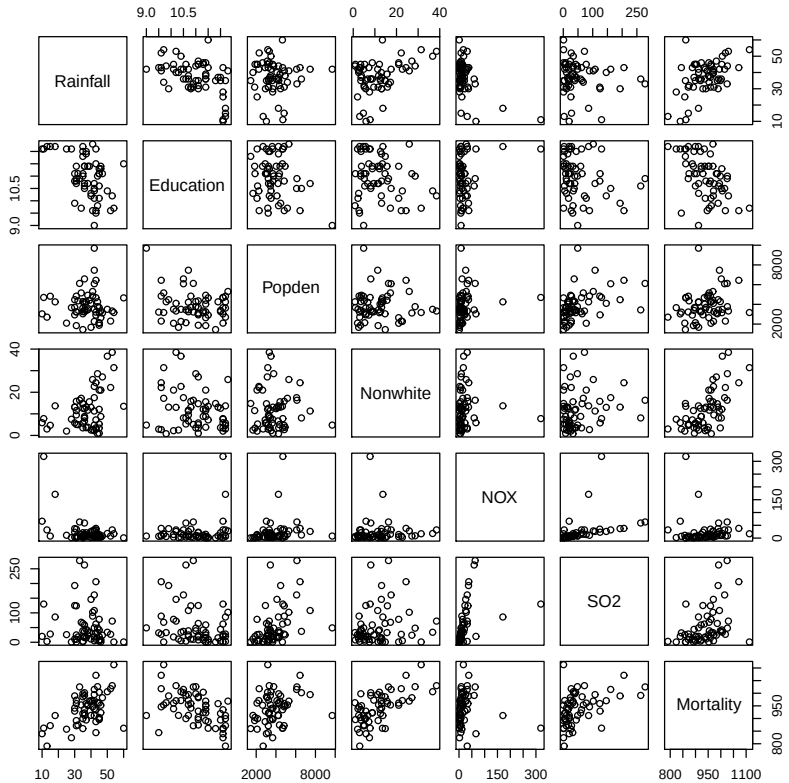
The radius of the circle is proportional to `Rainfall` (code in notes):



Graphics in R – more dimensions

Several variables (well, 7 is already quite tricky to analyse!)

```
pairs(airpoll)
```



Graphics in R

Even in 3 dimensions, taking all the 2-dimensional views of pairs of marginal distributions can miss interesting features.

This can even happen in 2 dimensions – think of an example when an outlier would be missed by looking at the marginal distributions (the distributions of each individual variable).

With 2-dimensional views of pairs of variables in more dimensions, more features may be missed. Nevertheless, since we are restricted to 2-dimensional views, these plots are about as good as we can get.

- Some packages allow 3-dimensional plots of data which can be interactively rotated (by e.g., moving the mouse).

Graphics in R

Summary

- R has a huge graphical capability. . .
- . . . but even this doesn’t help if there are lots of dimensions

So what can we do with a lot of dimensions?

Graphics in R – more dimensions

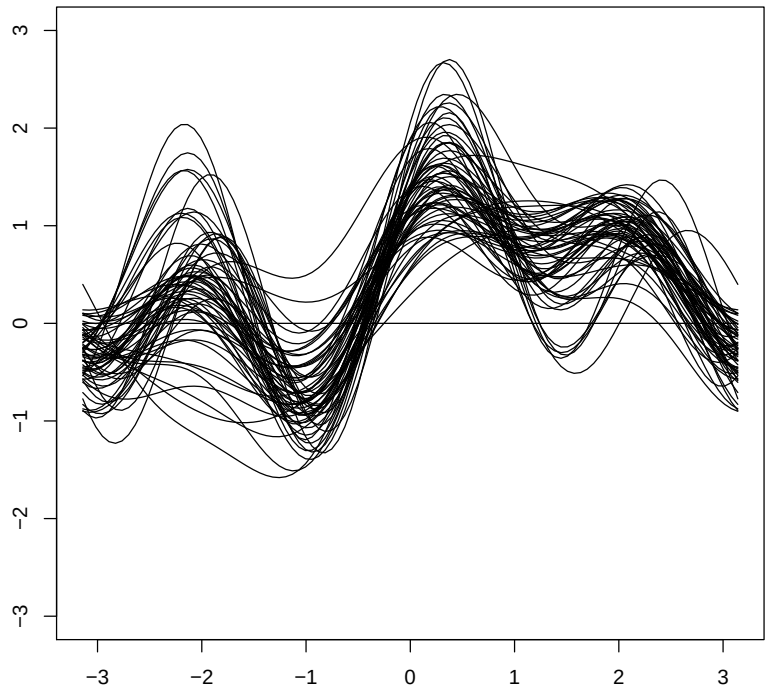
Andrews plots

- maps p-dimensional data $\{x_i\}$ onto 1-dimensional $\{f_i(t)\}$ for a variable t
- plotting $\{f_i(t)\}$ for (say) $t \in [-\pi, \pi]$ yields a 1-dimensional representation
- if the f_i are defined suitably (see notes), then many statistical properties are preserved
- available in various R packages (e.g., `andrews`)

Graphics in R – more dimensions

Andrews plots

```
andrews(airpoll,ymax=3)
```



Graphics in R – more dimensions

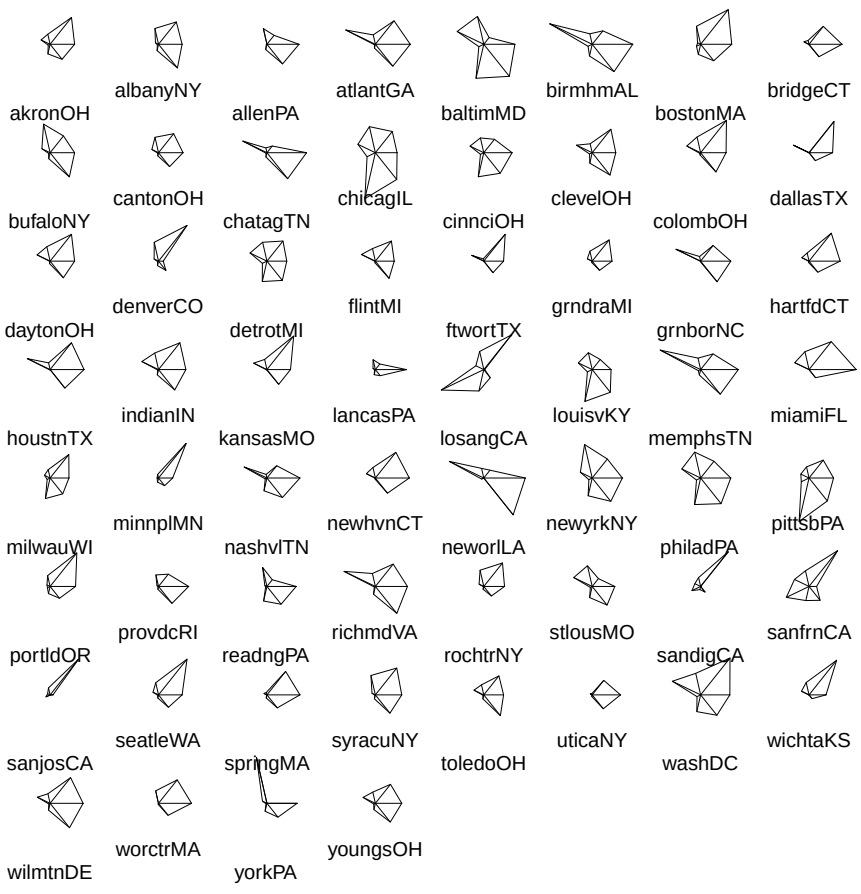
Star plots

- reasonable for *p* up to (maybe) 15, but not too many observations
- each observation is represented by a polygon
- value of each variable corresponds to length of the vector to each vertex
  - can compare individual points
  - may be able to identify similar points
- in `base` R package

Graphics in R – more dimensions

Star plots

```
star(airpoll)
```



Graphics in R – more dimensions

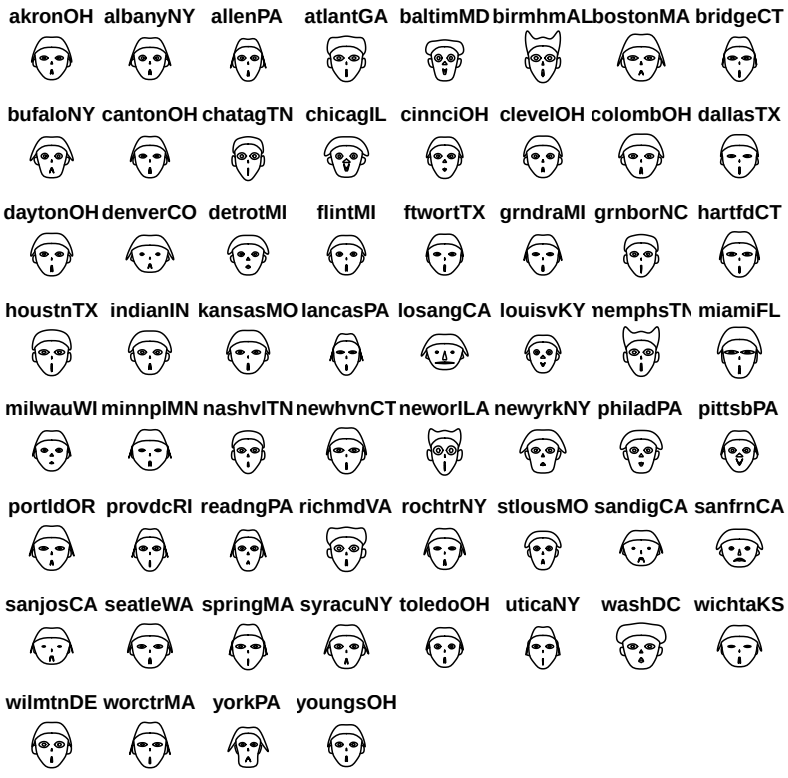
Chernoff faces

- reasonable for *p* up to (maybe) 15, but not too many observations
- each observation is represented by a face (or similar graphic)
- value of each variable corresponds to length/angle/curvature of a specific feature
  - can compare individual points
  - may be able to identify similar points
- in various R packages (e.g., `TeachingDemos`)

Graphics in R – more dimensions

Chernoff faces

```
faces(airpoll)
```



Graphics in R

Summary

- R is very good for graphics. . .
- . . . but visualisation of raw data is difficult to interpret in any moderate number of dimensions

We conclude that we need to do some pre-processing of the data before visualisation