CHAPTER 3
EXPLORATORY DATA ANALYSIS

Kwankamon Dittakan, Ph.D.
College of Computing
Prince of Songkla University
Phuket, Thaliand

CONTENT

1. Why Exloratory Data Analysis?
2. Summary Statistics
3. Visualization

EXPLORATORY DATA ANALYSIS

- **Exploratory data analysis** (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.
- Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.

FOUR KEY CONCEPTS IN EXPLORING DATA

The four R's of exploratory data analysis proposed by Velleman and Hoaglin

1. Revelation refers to data visualisation
2. Residual refers to the set of differences between the observed values of a variable and its predictions from some mathematical model.
3. Re-expression refers to the application of mathematical transformation to one or more variables.
4. Resistance refers to the ability of a data characterisation to avoid the undue influence of outliers to other data anomalies.

EXPLORING A NEW DATASET

A general strategy often uses for exploring a new data set

1. Assess the general characteristics of the dataset, e.g.:
   - How many records do we have? How many variables?
   - What are the variable names? Are they meaningful?
   - What types is each variable? <numeric, categorical, logical>
   - How many unique values does each variable have?
   - What value occurs most frequently, how often does it occur?
   - Are there missing observations? If so, how frequently does this occur?
2. Examine descriptive statistics for each variable

EXPLORING A NEW DATASET

A general strategy often uses for exploring a new data set (CONT.)

3. Where possible – certainly for any variable of particular interest- examine exploratory visualisations.
4. Again, where possible, apply the procedures to look for data anomalies
5. Look at the relations between key variables
6. Summarised these results in the form of a data dictionary, to serve as a basis for subsequent analysis and explanation of the results.

TECHNIQUES USED IN DATA EXPLORATION

In EDA, as originally defined by Tukey

- The focus was on visualization
- Clustering and anomaly detection were viewed as exploratory techniques
- In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory

In our discussion of data exploration, we focus on

1. Summary statistics
2. Visualization

SUMMARY STATISTICS

- The Chile data frame from the car package describes voter intention prior to an election in Chile in 1988.
- This data frame has 2700 records and 8 variables.

```
dim(Chile)

## [1] 2700    8
```

SUMMARY STATISTICS

- The summary of data frame is suggested to explore

```
BasicSummary(Chile)

##      variable    type levels topLevel topCount topFrac missFreq missFrac
## 1      region  factor      5       SA      960   0.356        0    0.000
## 2  population integer     10   250000     1300   0.481        0    0.000
## 3         sex  factor      2        F     1379   0.511        0    0.000
## 4         age integer     54       21       96   0.036        1    0.000
## 5   education  factor      4        S     1120   0.415       11    0.004
## 6      income integer      8    15000      768   0.284       98    0.036
## 7   statusquo numeric   2093 -1.29617      201   0.074       17    0.006
## 8        vote  factor      5        N      889   0.329      168    0.062
```

SUMMARY STATISTICS

Variable types:

- One of the key data characteristics included in the preliminary data summary just described is variable type.
- There are three different variable types: numerical, categorical, and ordinal.
- Numerical variable
  - Many data analysis problems are concerned with numerical data.
  - An important characteristics of numerical data is that we can apply many mathematical operations to it, computing sums, differences, products, quotients, averages, square roots and many other combination and/or transformation.
  - The basis for descriptive statistics such as mean and standard deviation also suggested.

SUMMARY STATISTICS

Variable types (CONT.)

- Nominal variable (Categorical variable)
  - None of the mathematical operations could be applied.
  - All we can do is count and compare, to answer the following question
    - How many distinct value or "Levels" does the variable exhibit?
    - How often does each of these levels occur in the dataset?
    - How des the behaviour of another variable X vary over the levels of C?
  - Some visualisation techniques can be applied

SUMMARY STATISTICS

Variable types (CONT.)

- Ordinal variable
  - Also referred to as ordered categorical variables or ordered factor.
  - These variables assume non-numeric values so the full rage of mathematical operations is not available to work with them
  - However, they do possess an inherent order, so we can say that one value of the variable is "smaller than" or "precedes" another value.
  - Example: "low, medium, and high" or survey results an order scale such as "strongly agree, agree, no opinion, disagree, strongly disagree".
  - We may compute medians or others characterisations the depend strictly on ordering

SUMMARY STATISTICS

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the
  data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode of a an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

SUMMARY STATISTICS

Percentiles

- For continuous data, the notion of a percentile is more useful.
- Given an ordinal or continuous attribute x and a number p between 0 and 100, the p^{th} percentile is a value x_p of x such that p% of the observed values of x are less than x_p.
- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.

SUMMARY STATISTICS

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

SUMMARY STATISTICS

Measures of Spread: Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- standard_deviation(x)= s_x
- However, this is also sensitive to outliers, so that other measures are often used.

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

(Mean Absolute Deviation) [Han]
(Absolute Average Deviation) [Tan]

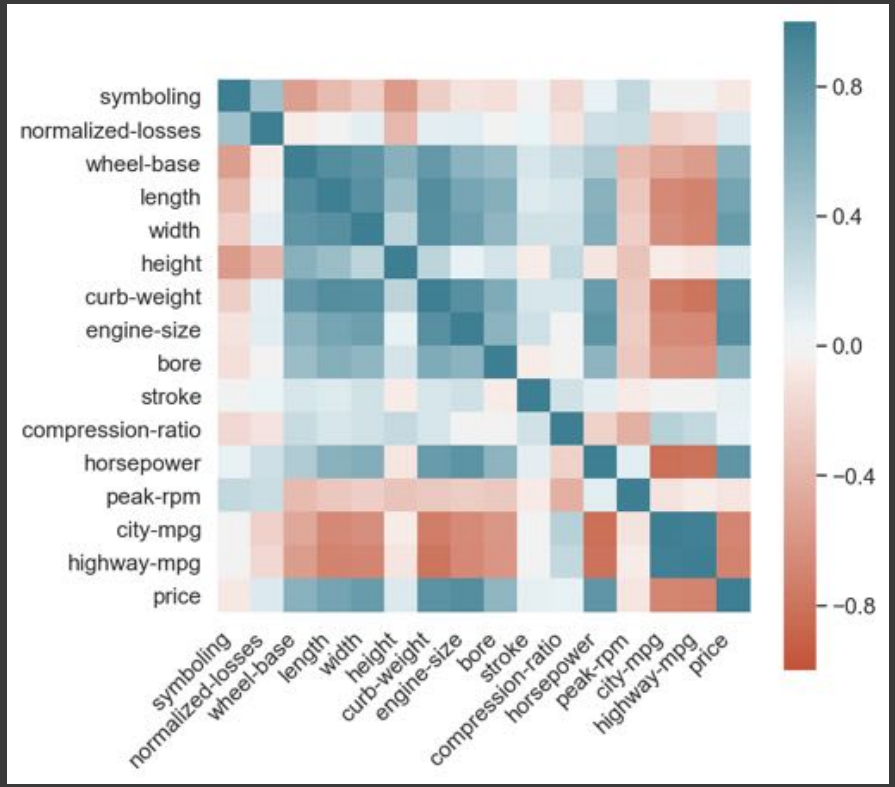
$$\text{MAD}(x) = median\left(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\right)$$

(Median Absolute Deviation)

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

SUMMARY STATISTICS

Correlation

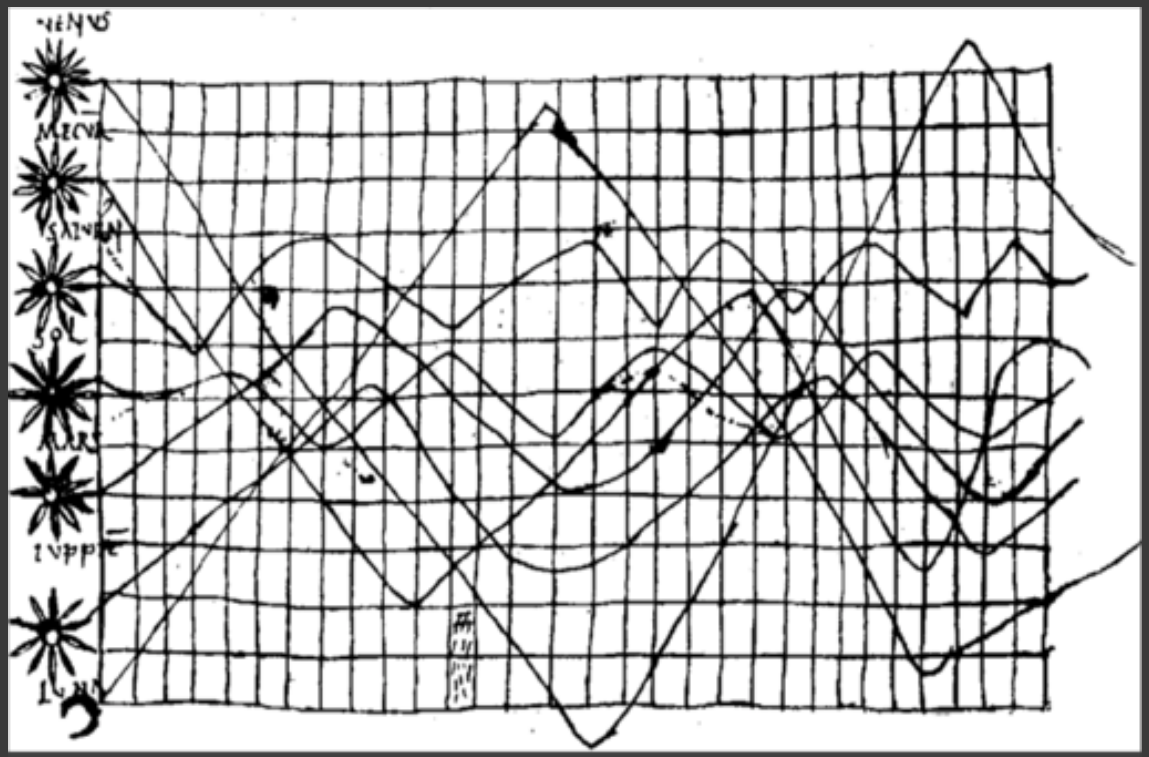


VISUALISATION

- Visualisation is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
- Visualisation of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

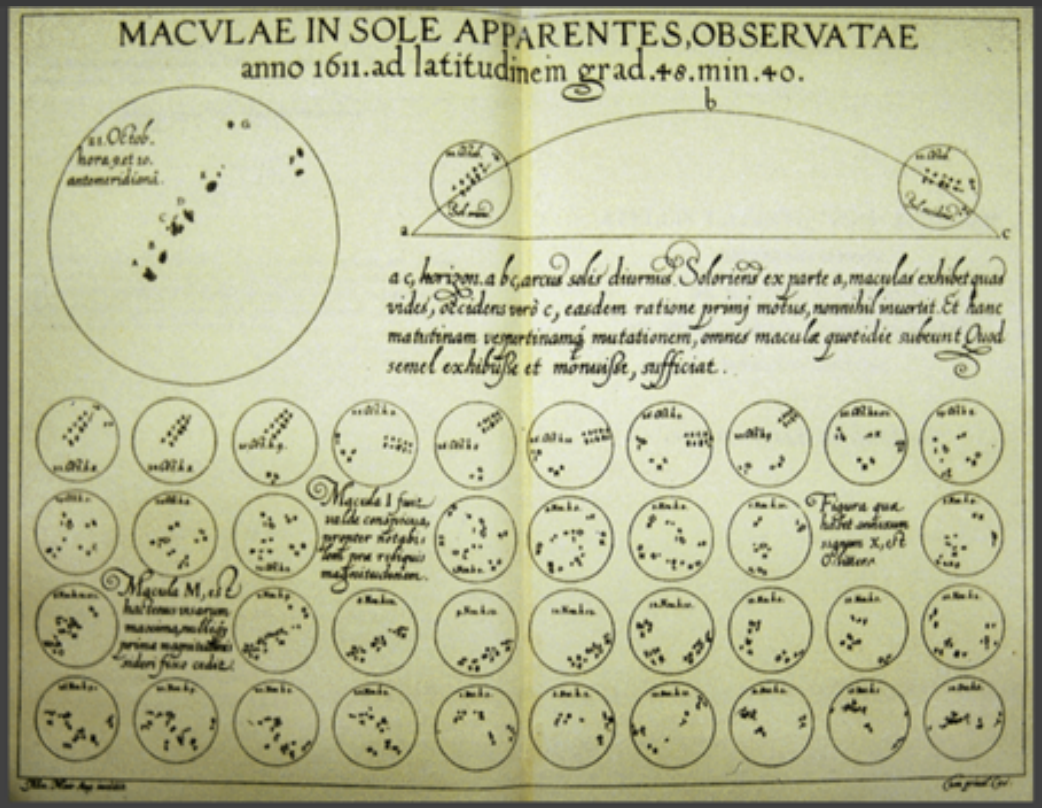
VISUALISATION IN HISTORY

VISUALISATION IN HISTORY



~950 AD Position of Sun, Moon and Planets

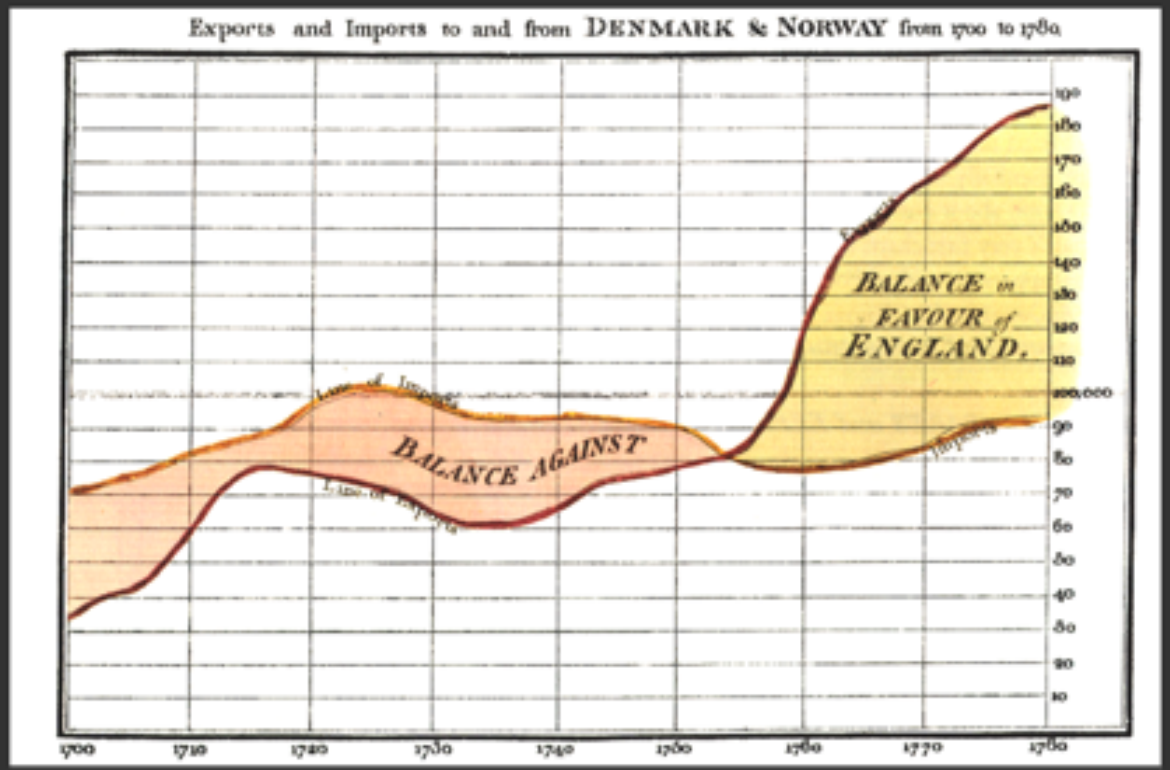
VISUALISATION IN HISTORY



0 BC

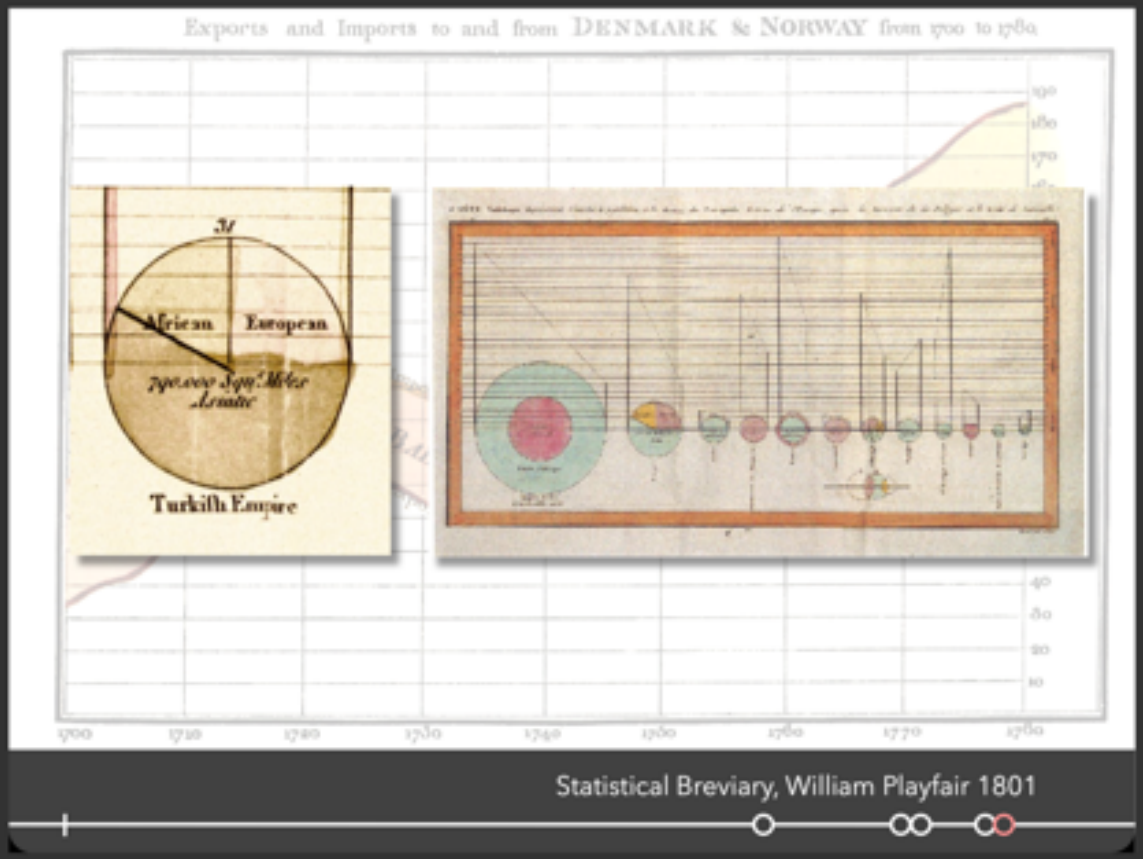
Sunspots over time, Scheiner 1626

VISUALISATION IN HISTORY



The Commercial and Political Atlas, William Playfair 1786

VISUALISATION IN HISTORY

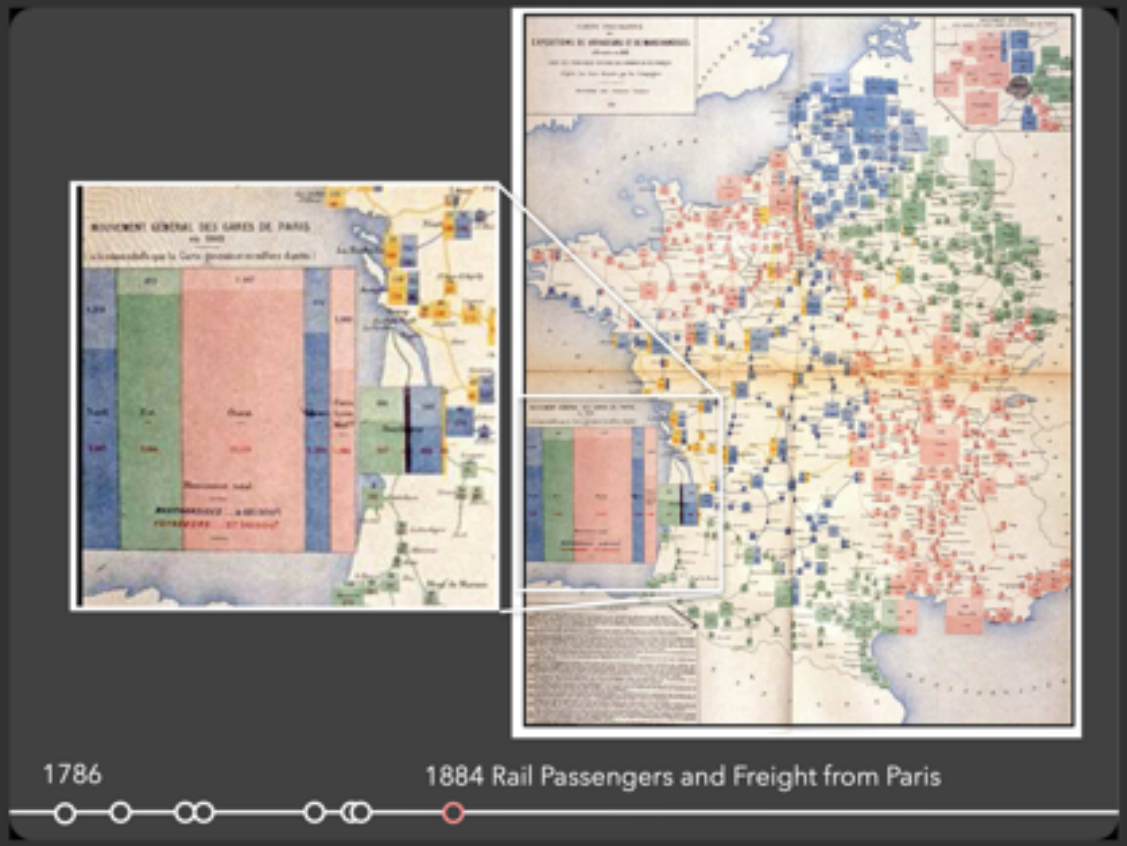


Statistical Breviary, William Playfair 1801

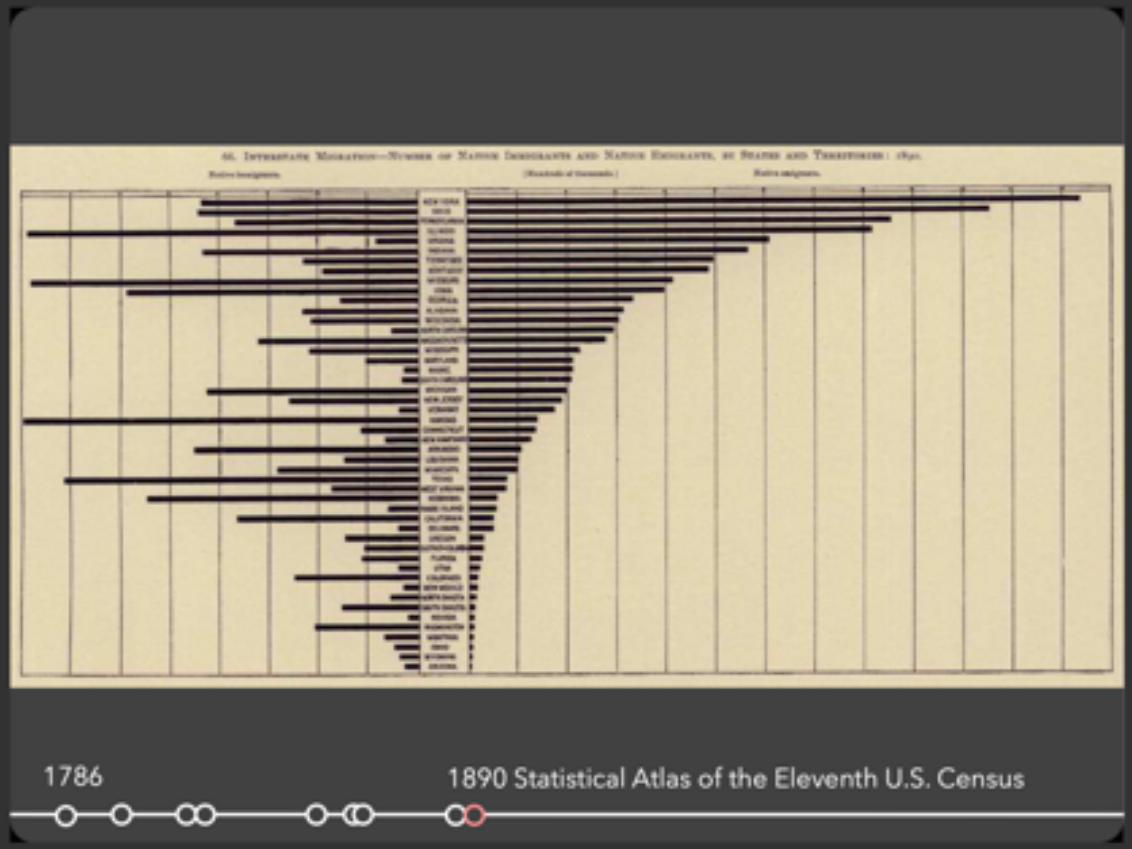
VISUALISATION IN HISTORY

1786 — 1864 British Coal Exports, Charles Minard

VISUALISATION IN HISTORY



VISUALISATION IN HISTORY



DATA VISUALISATION

Representation

- Is the mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

DATA VISUALISATION

Arrangement

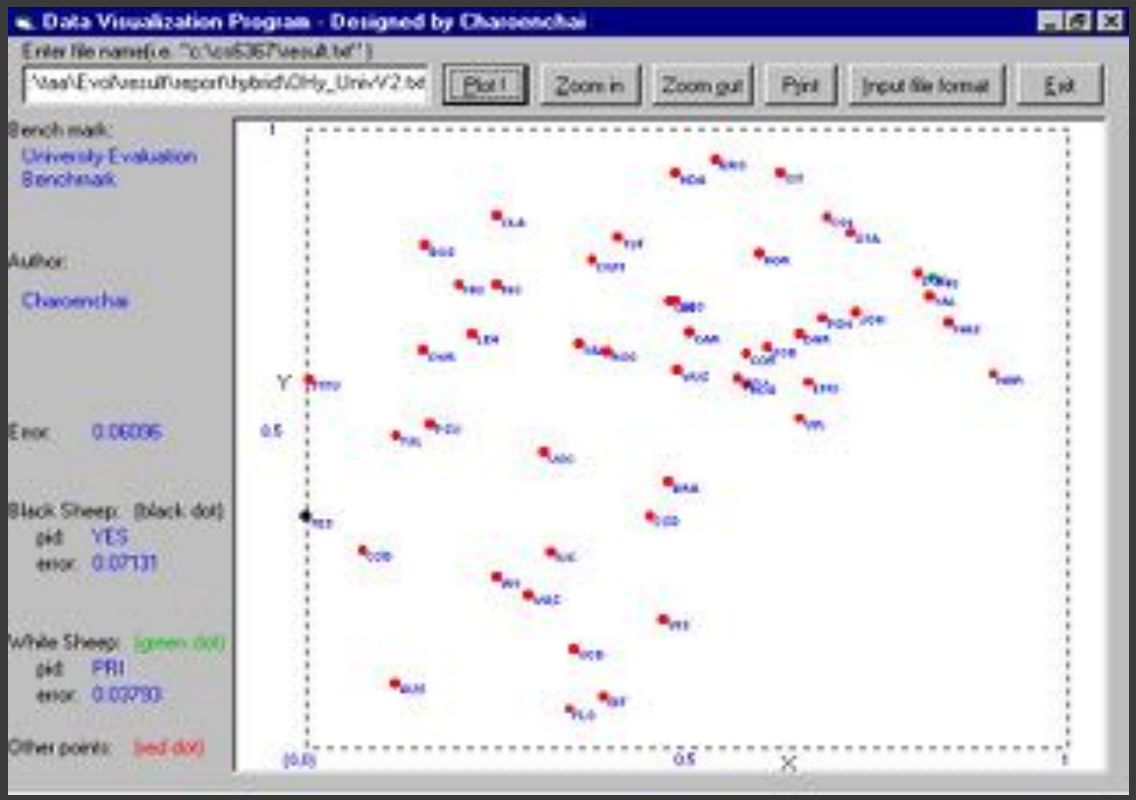
- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

DATA VISUALISATION

Example: Visualising Universities



DATA VISUALISATION

Selection

- Is the elimination or the de-emphasis of certain objects and attributes
- Selection may involve the choosing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be considered
- Selection may also involve choosing a subset of objects
  - A region of the screen can only show so many points
  - Can sample, but want to preserve points in sparse areas

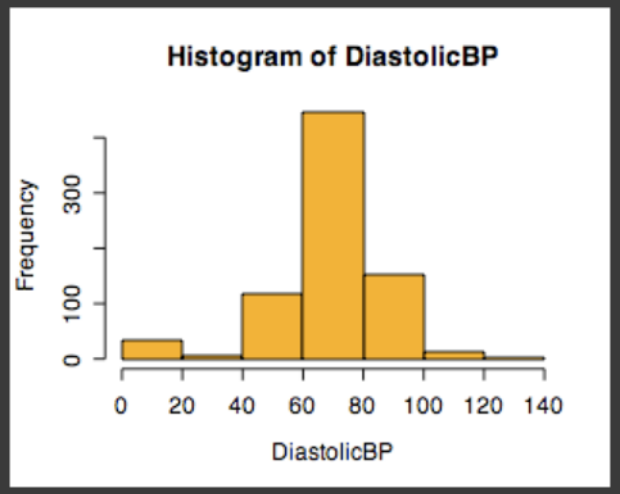
DATA VISUALIZATION TECHNIQUES

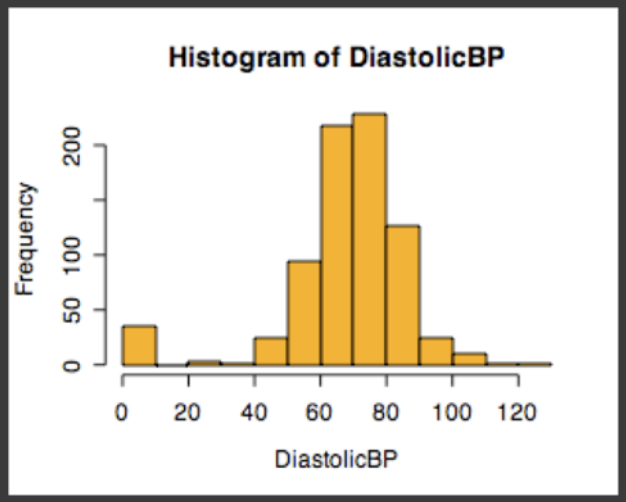
- One variable visualisation
- Two variable visualisation
- More than two variable visualisation
- Other visualization techniques

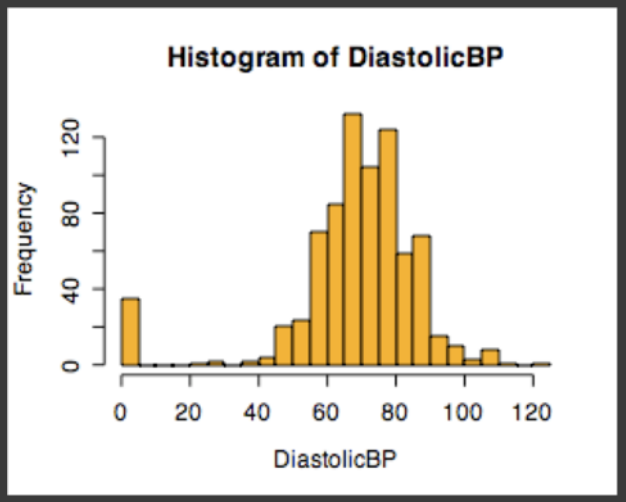
SINGLE VARIABLE VISUALIZATION

Histogram:

- Shows center, variability, skewness, modality,
- outliers, or strange patterns.
- Bin width and position matter
- Beware of real zeros







SINGLE VARIABLE VISUALIZATION

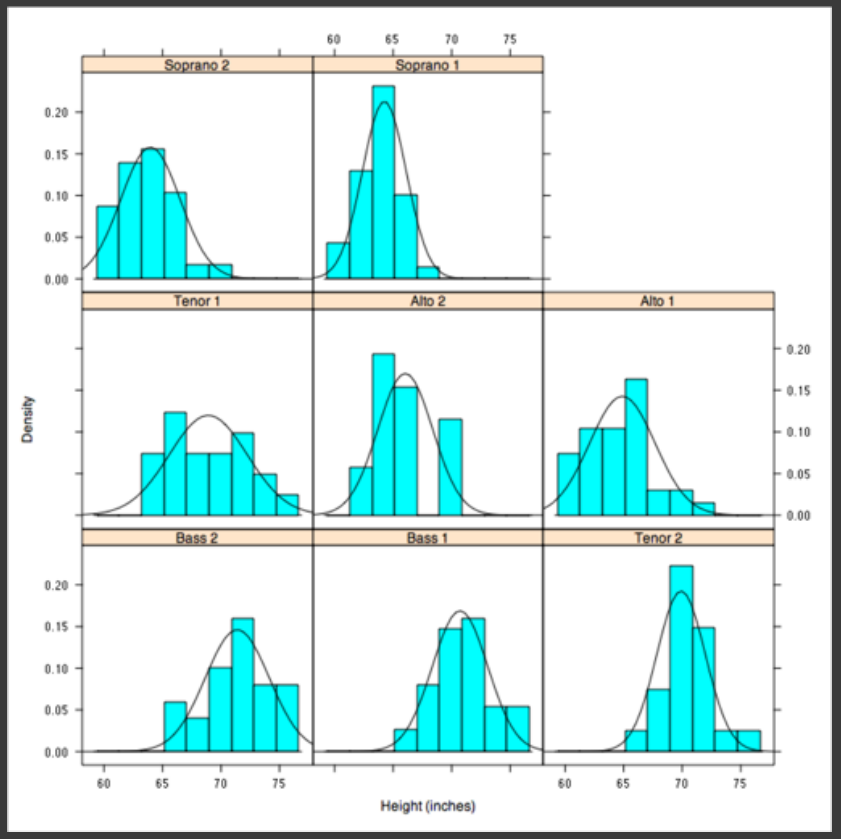
Histogram: Some issues with Histograms

- For small data sets, histograms can be misleading.
  - Small changes in the data, bins, or anchor can deceive
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.
- Histograms effectively only work with 1 variable at a time
  - But 'small multiples' can be effective

SINGLE VARIABLE VISUALIZATION

Histogram:

But be careful with axes and scales!



SINGLE VARIABLE VISUALIZATION

Histogram:

Smoothed Histograms - Density Estimates

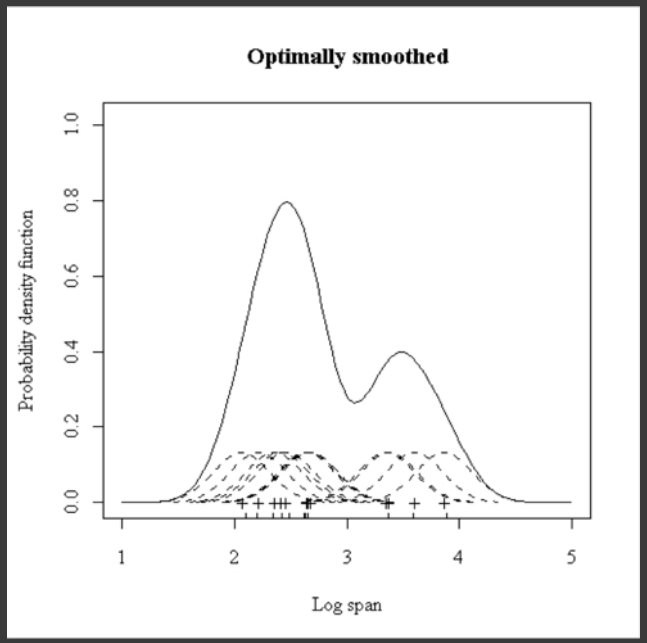
- Kernel estimates smooth out the contribution of each data point over a local neighborhood of that point.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

h is the kernel width

- Gaussian kernel is common:

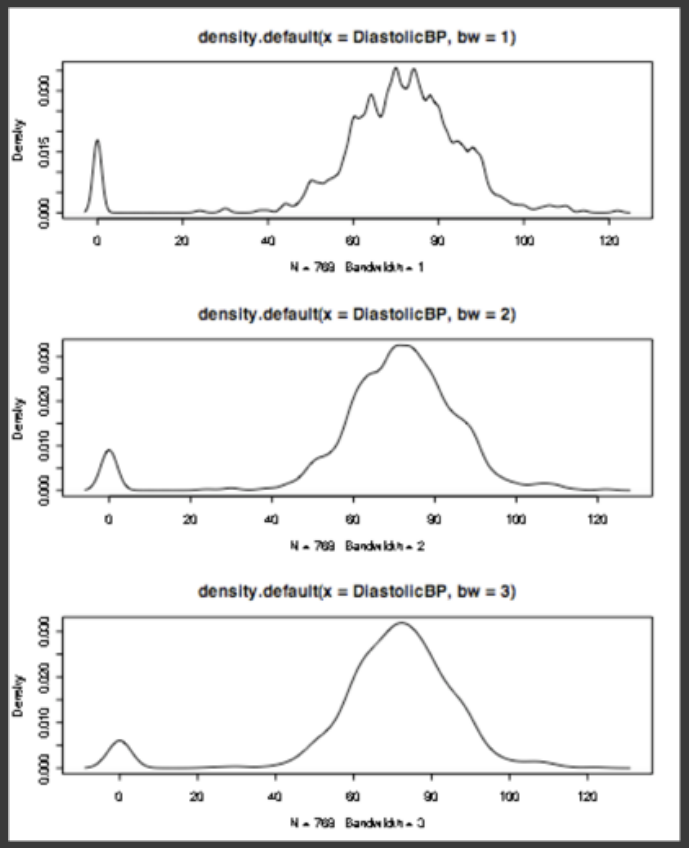
$$Ce^{-\frac{1}{2}\left(\frac{x-x(i)}{h}\right)^2}$$



SINGLE VARIABLE VISUALIZATION

Histogram:

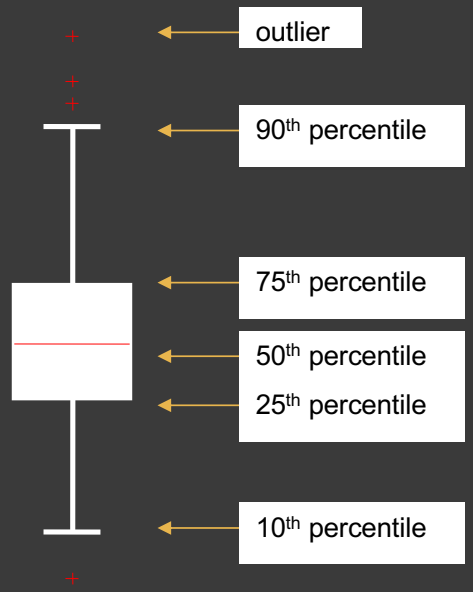
- Bandwidth choice is an art
- Usually want to try several



SINGLE VARIABLE VISUALIZATION

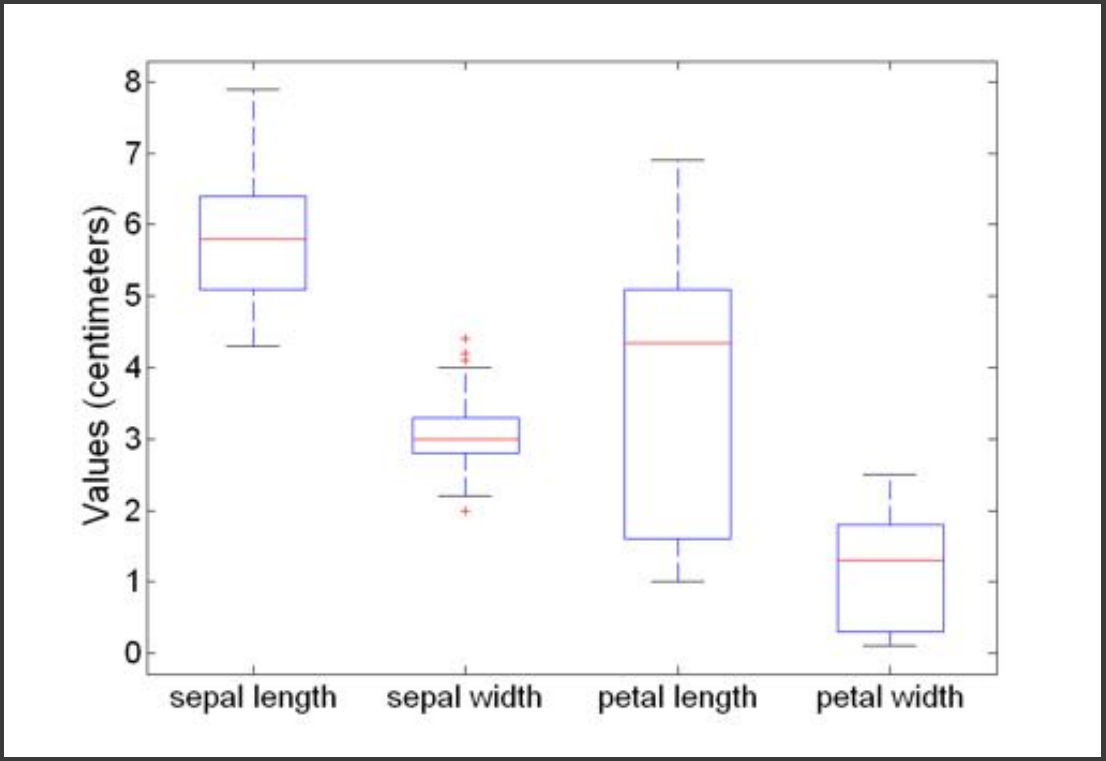
Boxplots

- Shows a lot of information about a variable in one plot
  - Median
  - IQR
  - Outliers
  - Range
  - Skewness
- Negatives
  - Overplotting
  - Hard to tell distributional shape
  - no standard implementation in software (many options for whiskers, outliers)



SINGLE VARIABLE VISUALIZATION

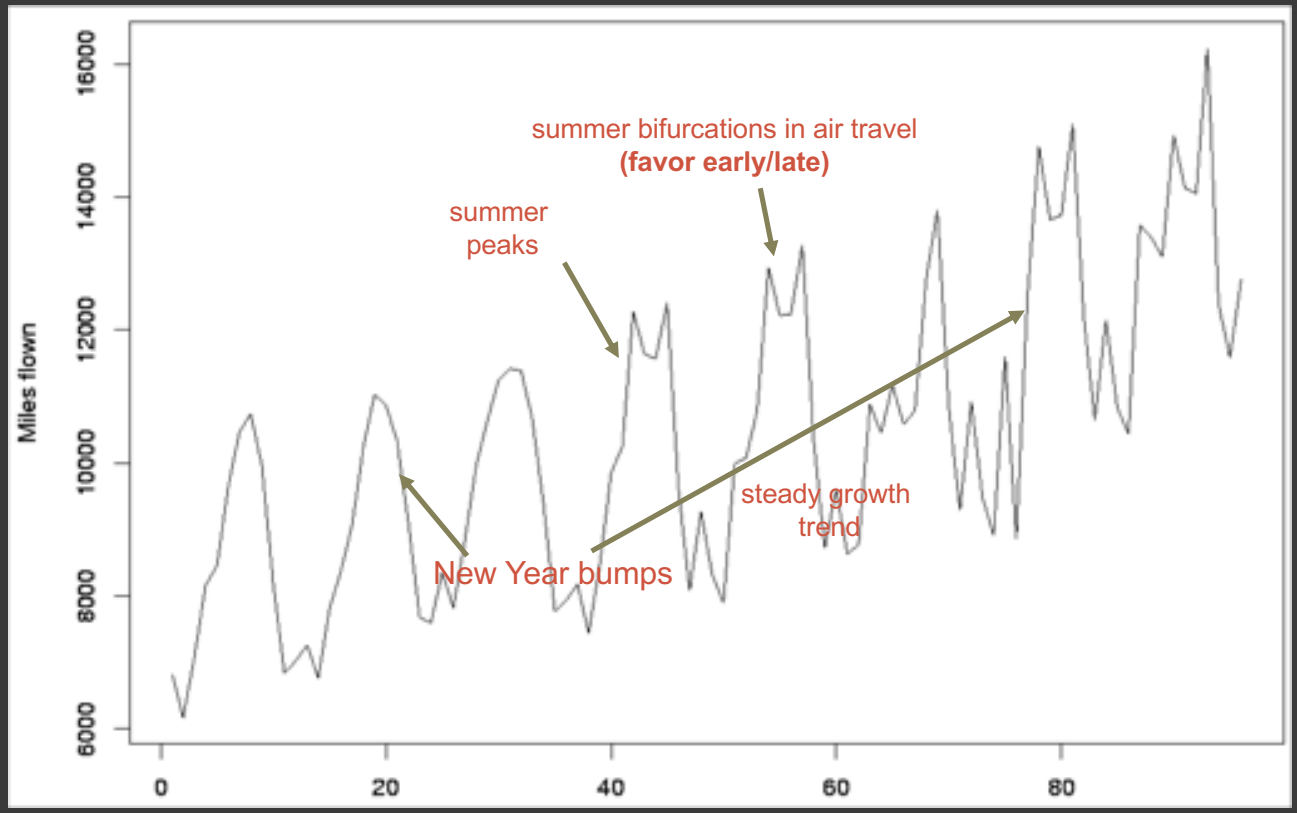
- Example of Box Plots
- Box plots can be used to compare attributes



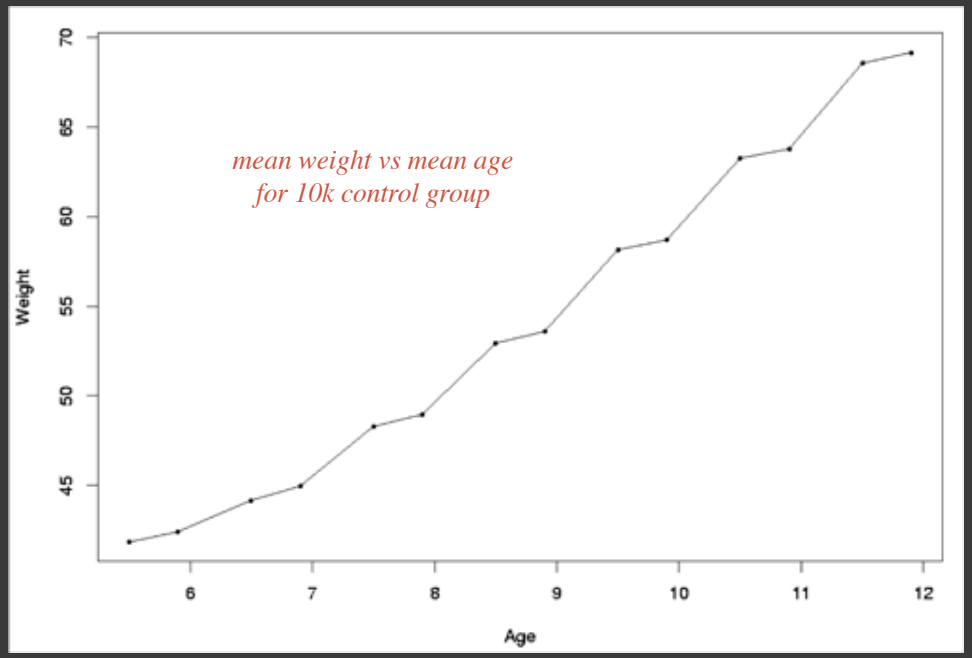
SINGLE VARIABLE VISUALIZATION

Time Series

If your data has a temporal component, be sure to exploit it



SINGLE VARIABLE VISUALIZATION



Time- Series Example

Scotland experiment:
"↑ milk in kid diet → better health" ?

20,000 kids:
5k raw, 5k pasteurize,
10k control (no supplement)

Would expect smooth weight growth plot.

Visually reveals
unexpected pattern (*steps*),
not apparent from raw data table.

Possible explanations:

Grow less early in year than later?

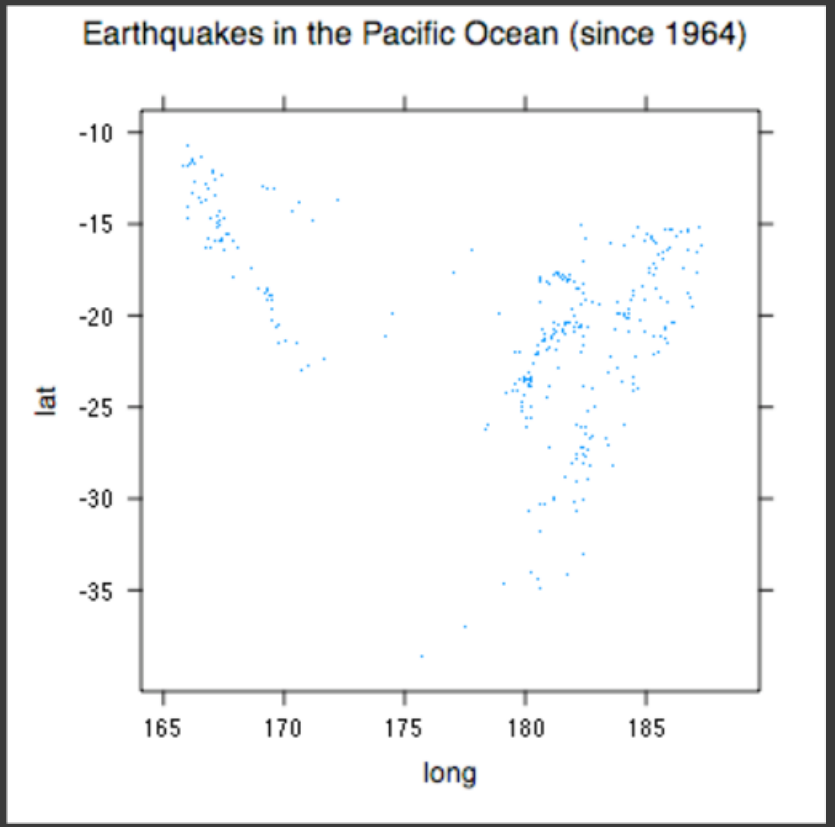
No steps in height plots; so why height ↑ uniformly, weight ↑ spurts?

Kids weighed in clothes: summer garb lighter than winter?

SINGLE VARIABLE VISUALIZATION

Spatial Data

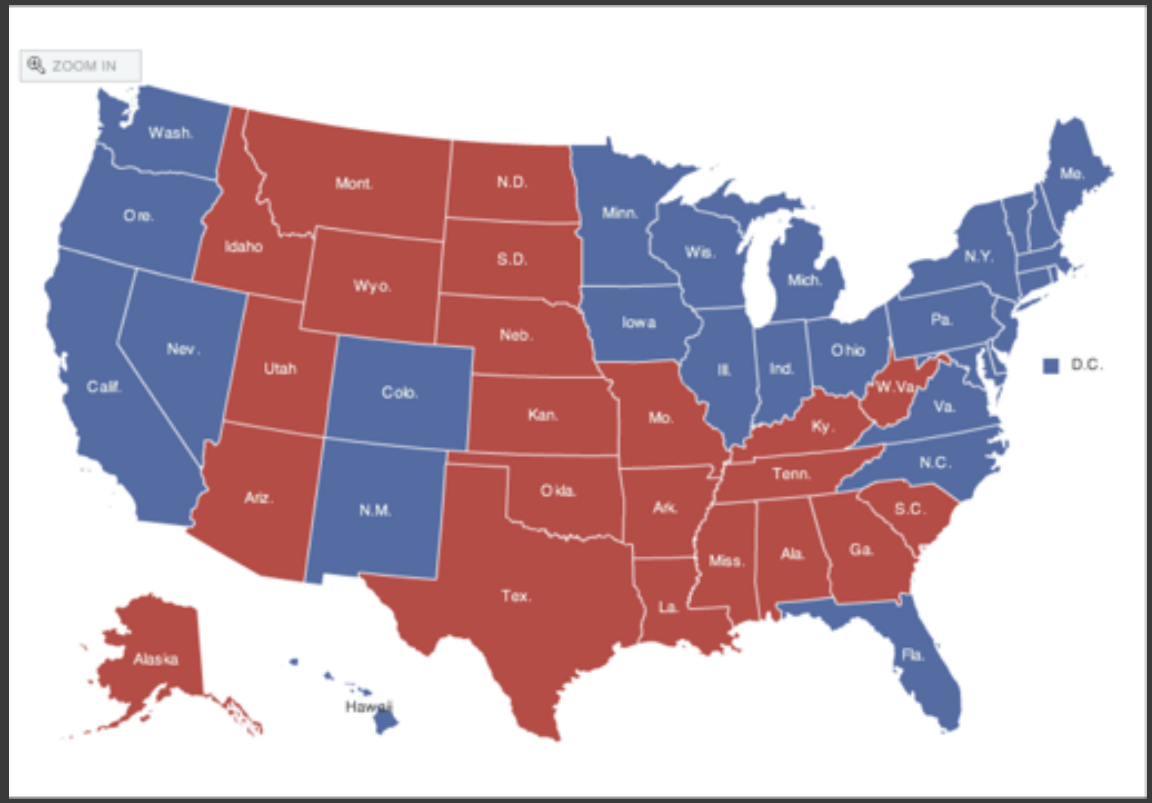
- If your data has a geographic component, be sure to exploit it
- Data from cities/states/zip cods – easy to get lat/long
- Can plot as scatterplot



SINGLE VARIABLE VISUALIZATION

Spatial data: choropleth Maps

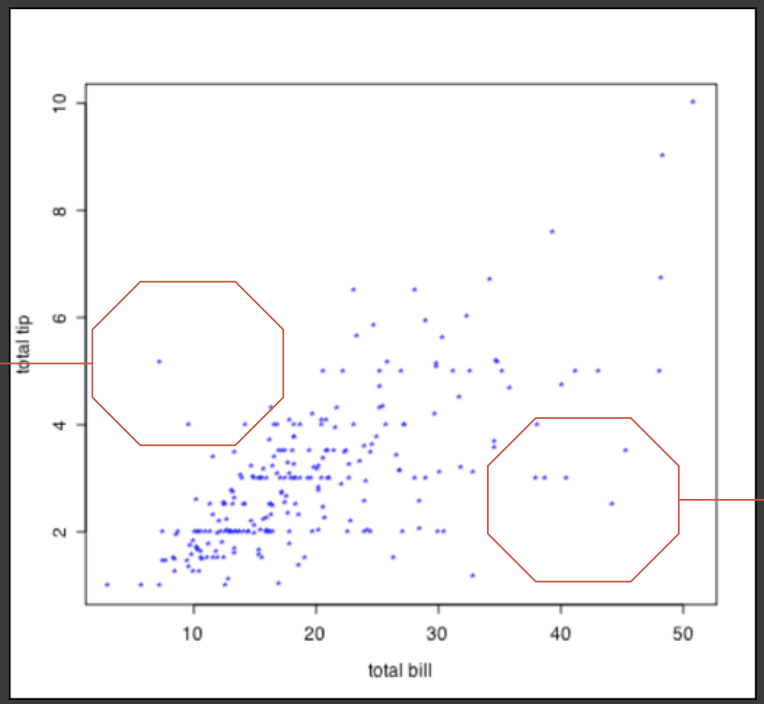
- Maps using color shadings to represent numerical values are called chloropleth maps
- http://elections.nytimes.com/2008/results/president/map.html



TWO VARIABLES VISUALISATION

- For two numeric variables, the scatterplot is the obvious choice

interesting?



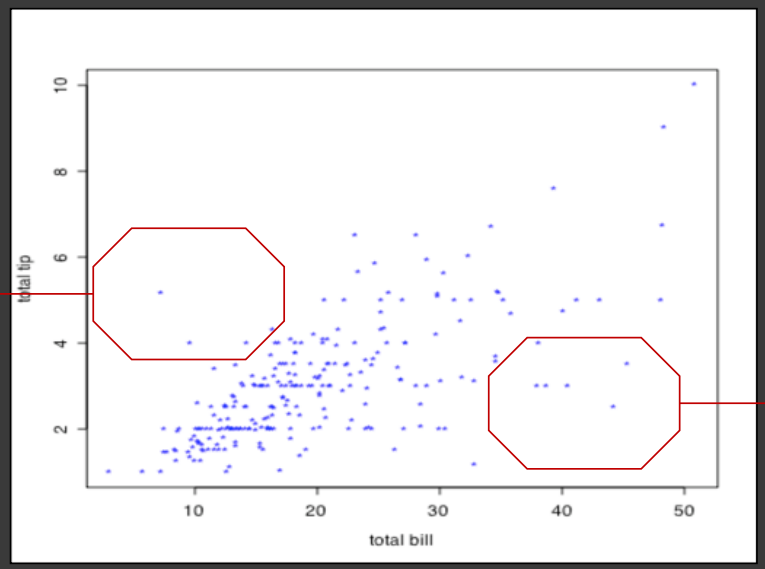
interesting?

TWO VARIABLES VISUALISATION

2D Scatterplots

- standard tool to display relation between 2 variables
- e.g. y-axis = response, x-axis = suspected indicator

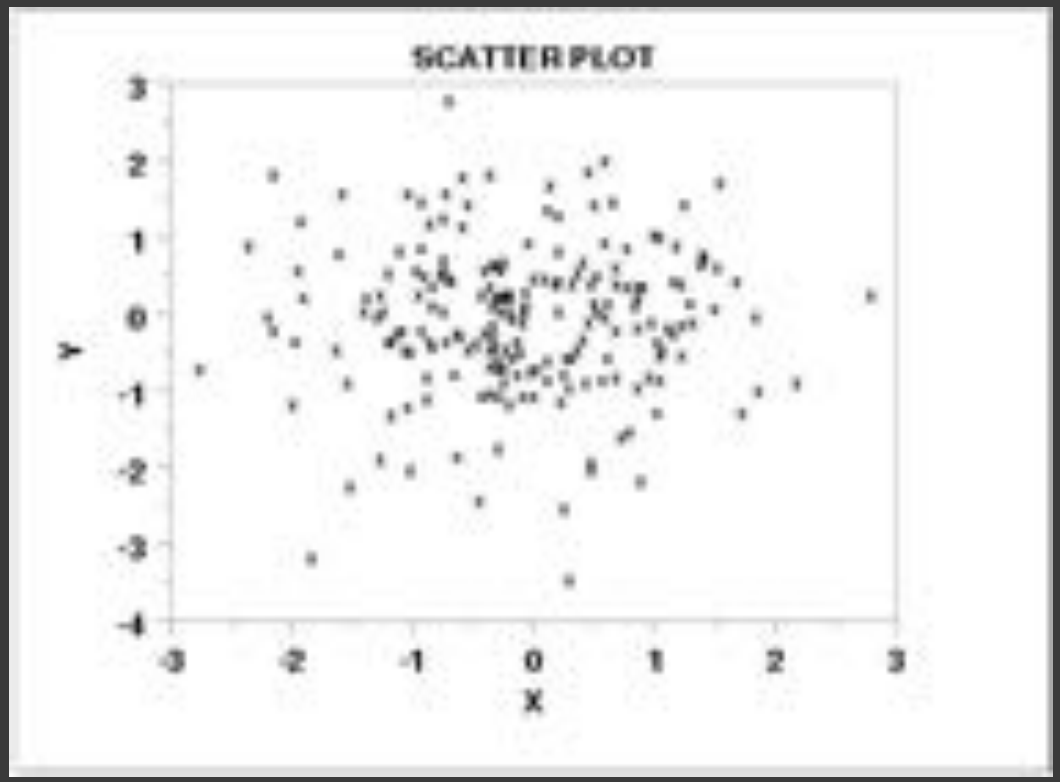
interesting?



- useful to answer:
  - x,y related?
    - linear
    - quadratic
    - other
  - variance(y) depend on x?
  - outliers present?

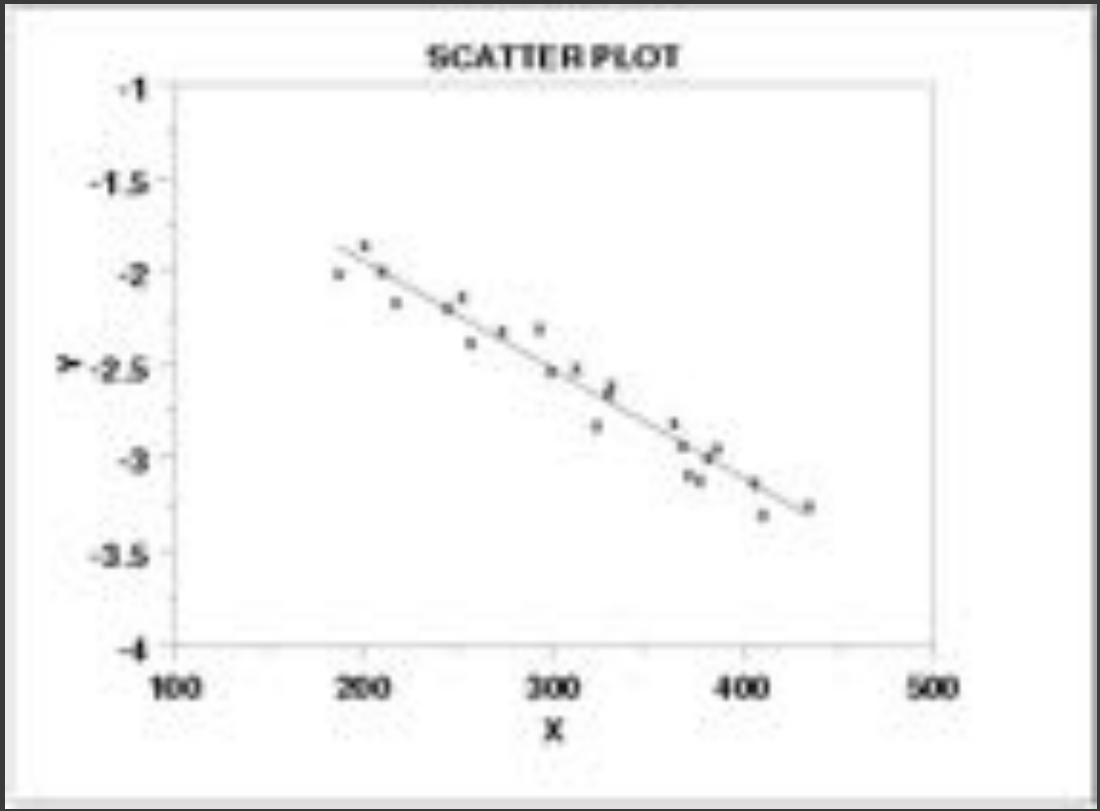
interesting?

TWO VARIABLES VISUALISATION



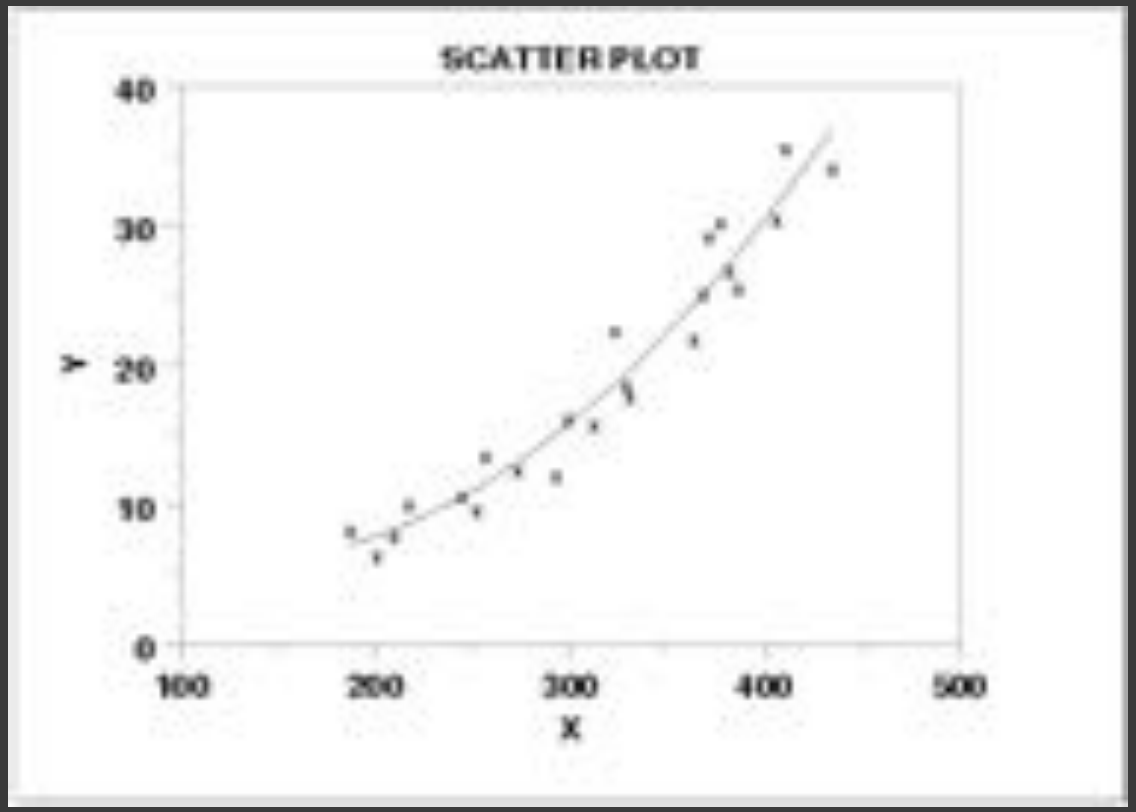
Scatter Plot: No apparent relationship

TWO VARIABLES VISUALISATION



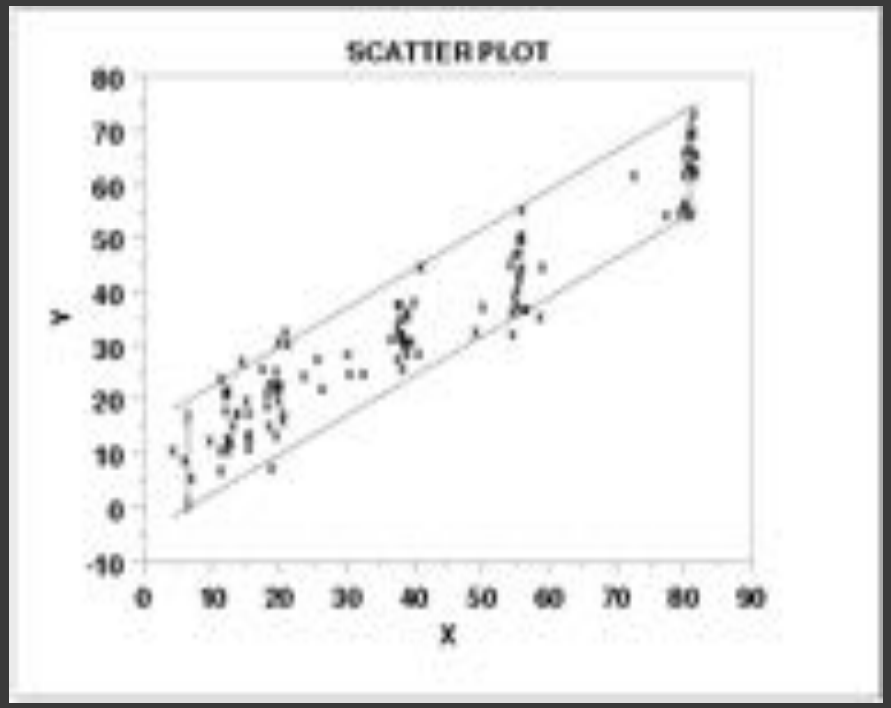
Scatter Plot: Linear relationship

TWO VARIABLES VISUALISATION



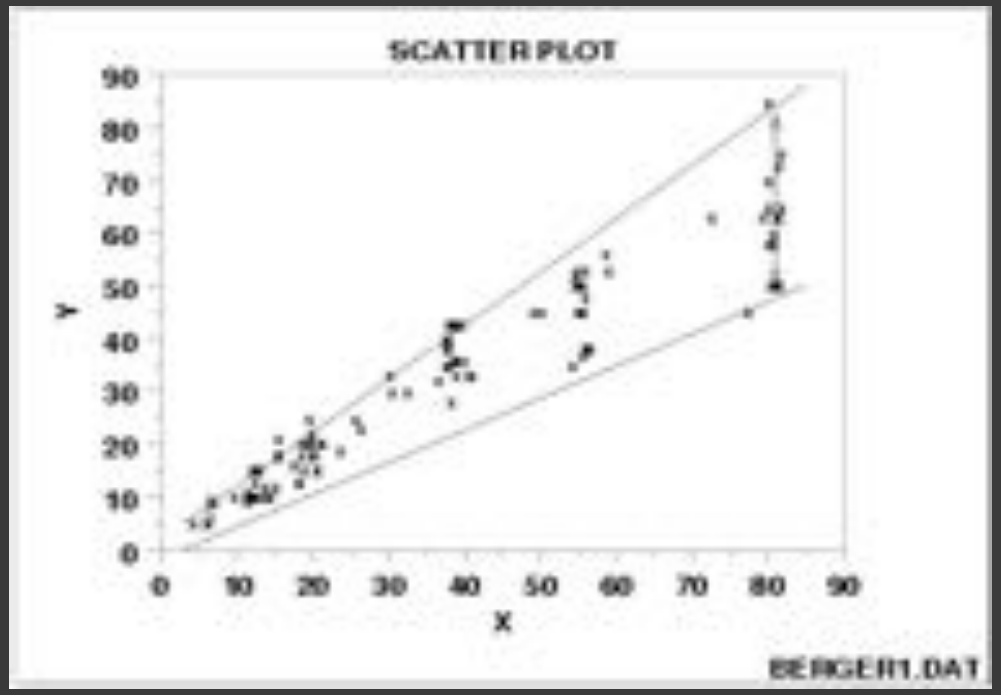
Scatter Plot: Quadratic relationship

TWO VARIABLES VISUALISATION



Scatter plot: Homoscedastic

TWO VARIABLES VISUALISATION



Scatter plot: Heteroscedastic

variation in *Y* differs depending on the value of *X*
e.g., Y = annual tax paid, X = income

TWO VARIABLES VISUALISATION

Scatterplots

- But can be bad with lots of data

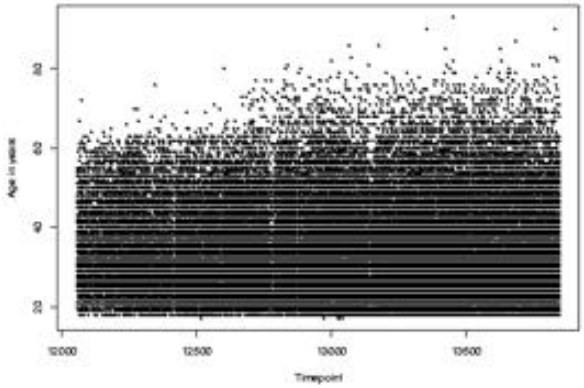
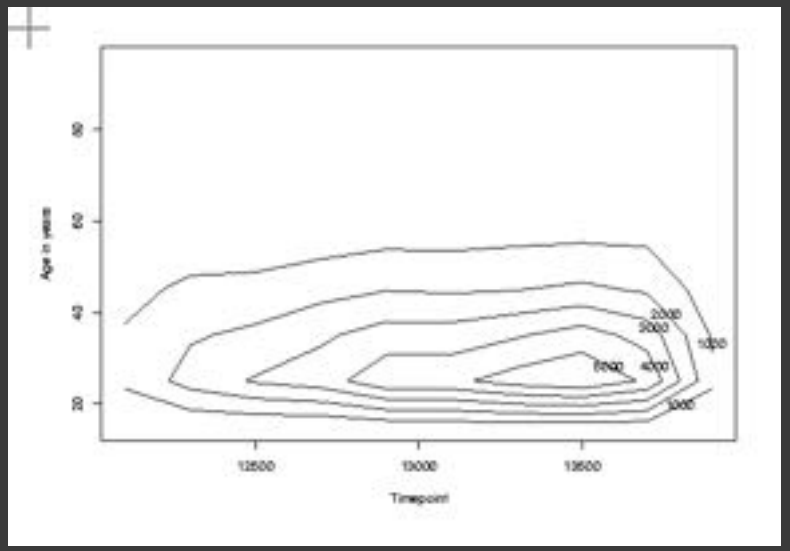


Figure 3.7: A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.

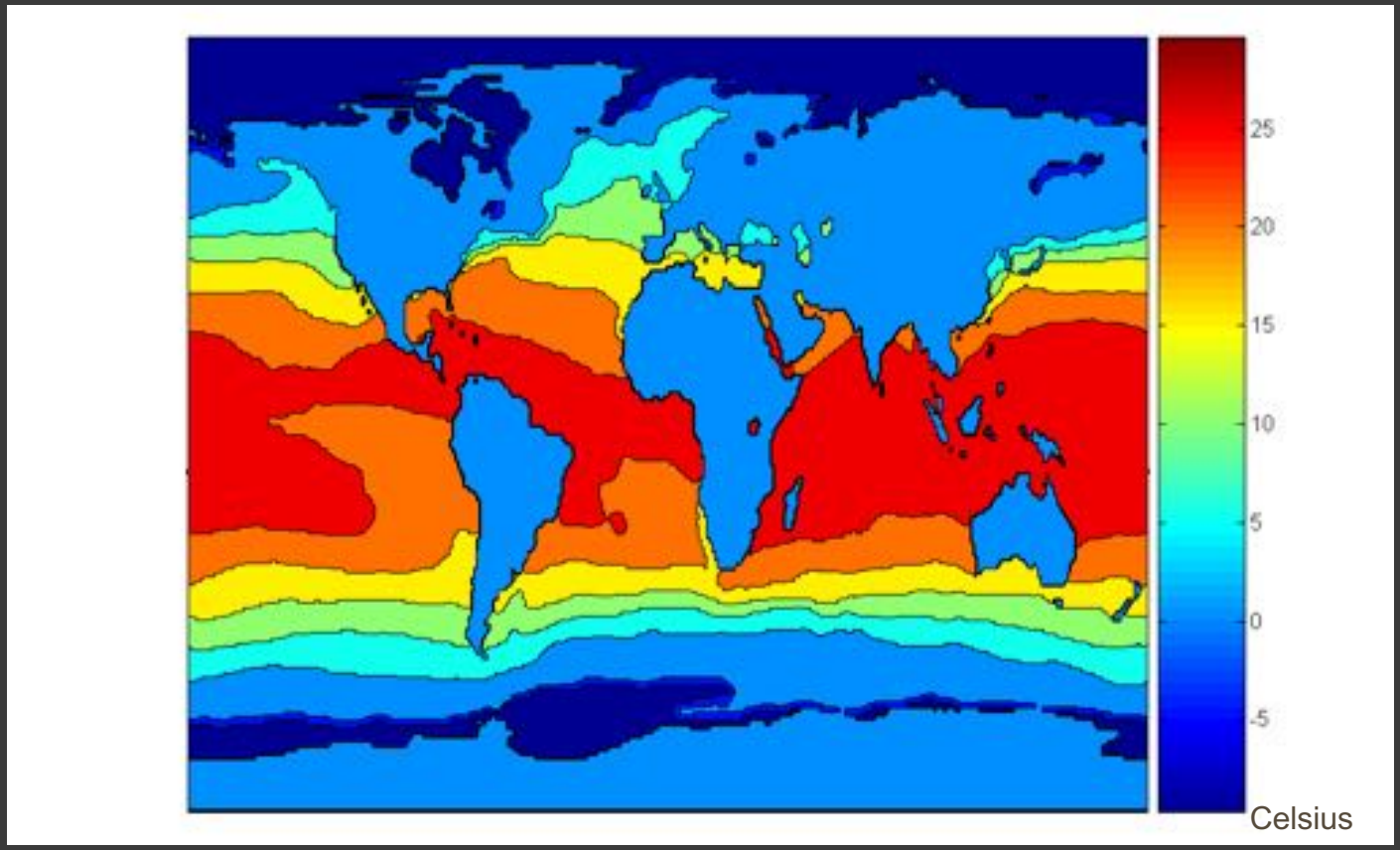
TWO VARIABLES VISUALISATION

What to do for large data sets? Contour plots

- Useful when a continuous attribute is measured on a spatial grid
- They partition the plane into regions of similar values
- The contour lines that form the boundaries of these regions connect points with equal values
- The most common example is contour maps of elevation
- Can also display temperature, rainfall, air pressure, etc.
  - An example for Sea Surface Temperature (SST) is provided on the next slide



TWO VARIABLES VISUALISATION

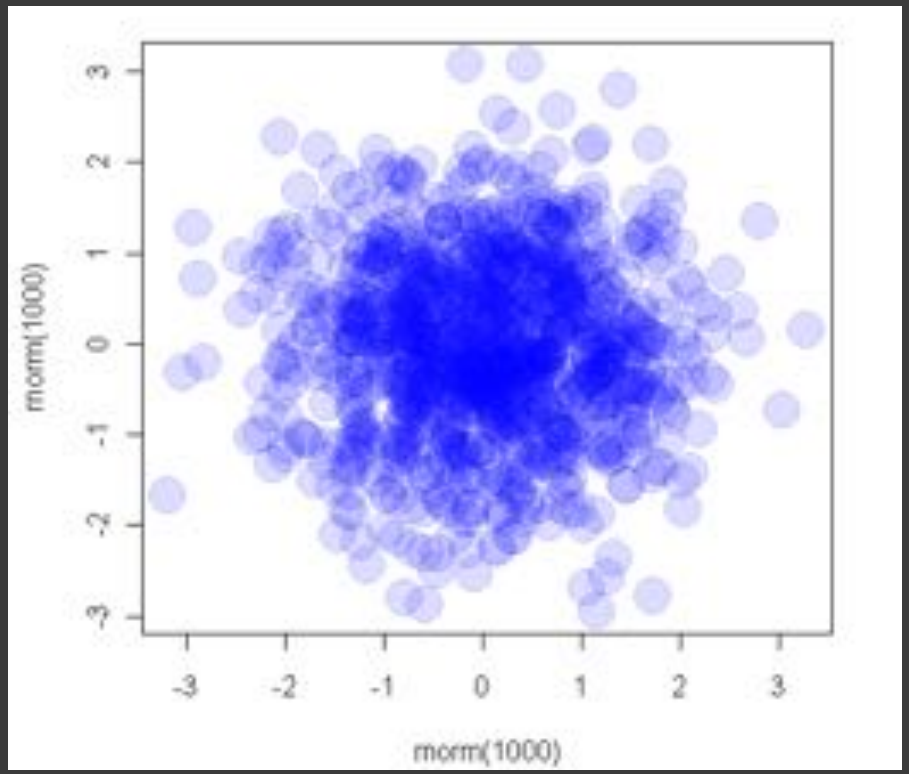


Contour Plot Example: SST
Dec, 1998

TWO VARIABLES VISUALIZATION

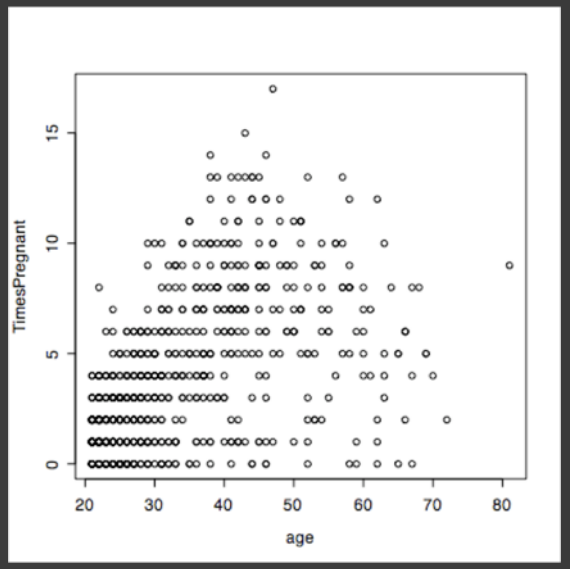
Transparent plotting

- Alpha-blending:
- plot(rnorm(1000), rnorm(1000), col="#0000ff22", pch=16,cex=3)

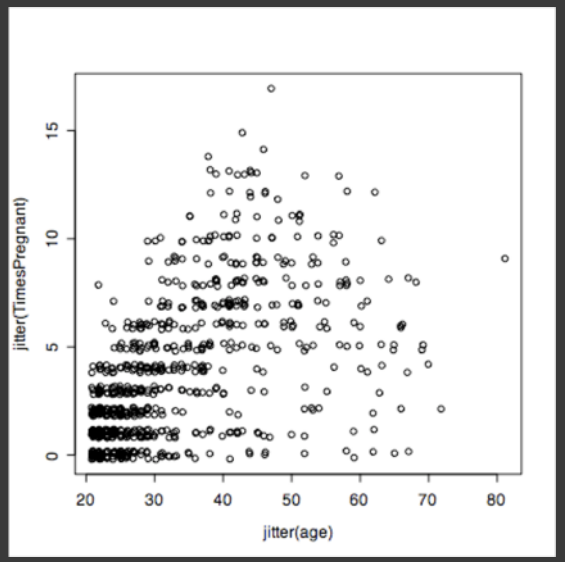


TWO VARIABLES VISUALISATION

Jittering points



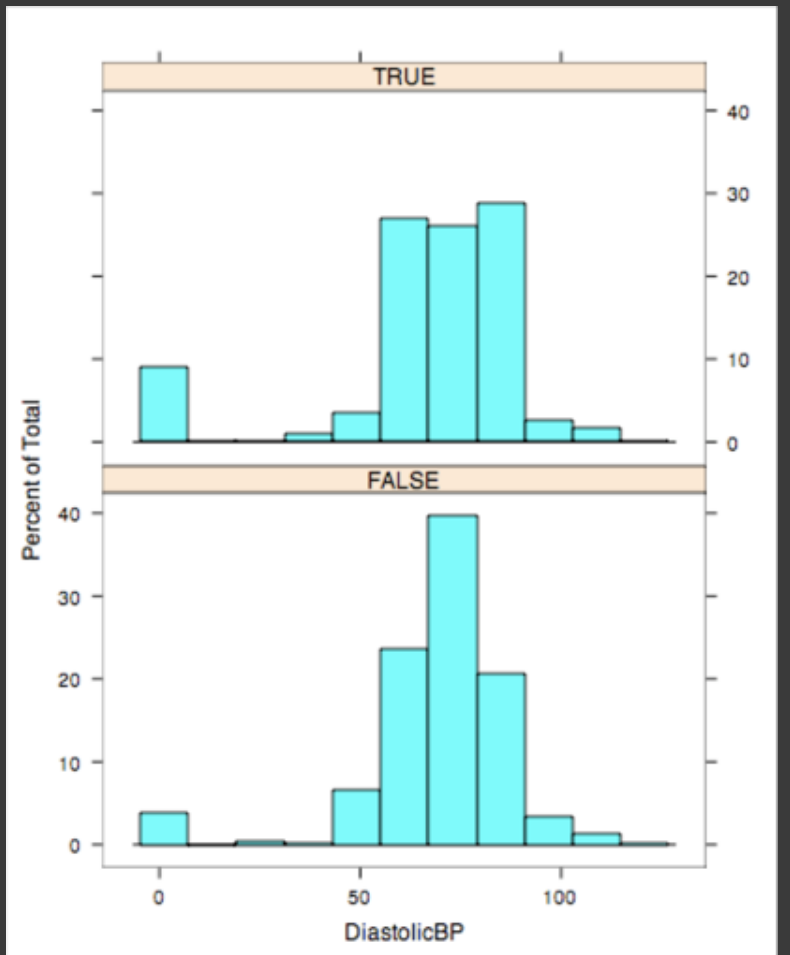
plot(age, TimesPregnant)



plot(jitter(age),jitter(TimesPregnant)

TWO VARIABLES VISUALISATION

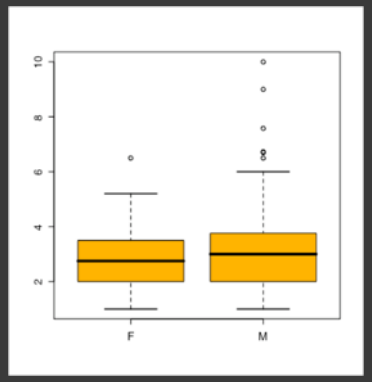
- If one variable is categorical, use small multiples
- Many software packages have this implemented as 'lattice' or 'trellis' packages

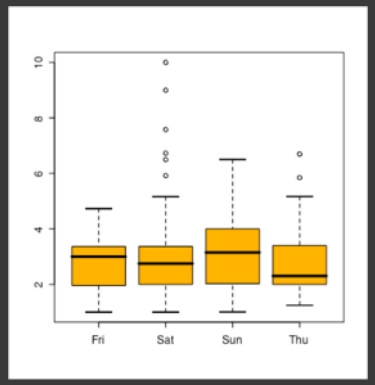


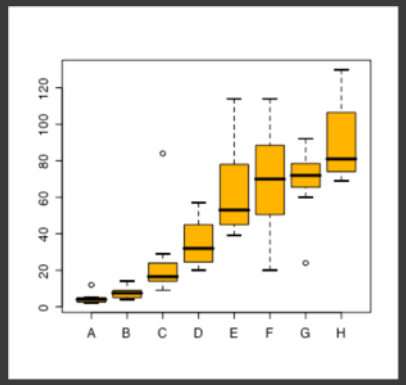

```
library('lattice')
histogram(~DiastolicBP | TimesPregnant==0)
```

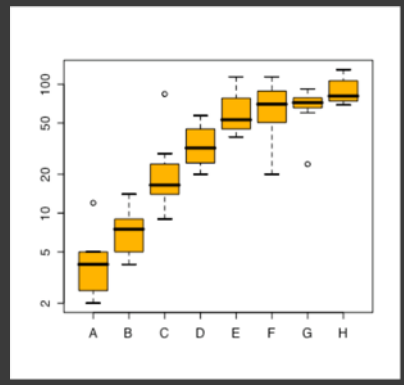
TWO VARIABLES VISUALISATION: ONE CATEGORICAL

- Side by side boxplots are very effective in showing differences in a quantitative variable across factor levels
  - tips data
    - do men or women tip better
  - orchard sprays
    - measuring potency of various orchard sprays in repelling honeybees







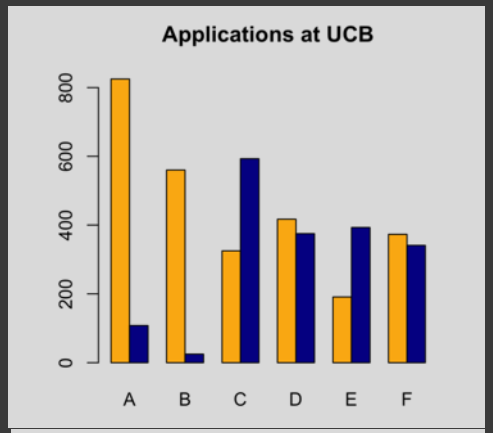


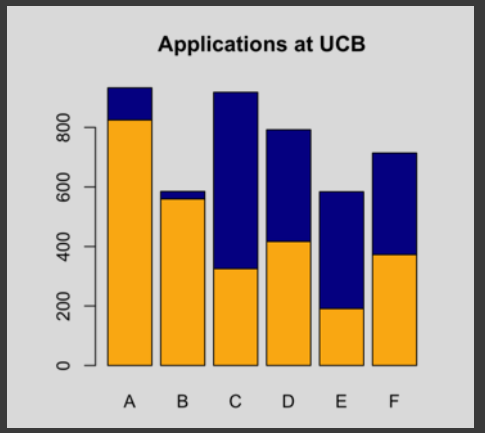
TWO VARIABLES VISUALISATION: ONE CATEGORICAL

Barcharts and Spineplots

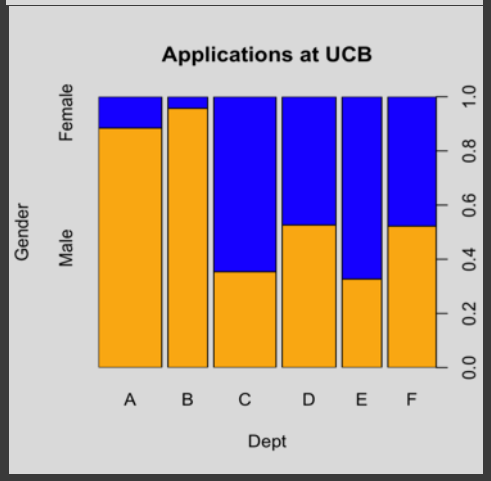
- *stacked barcharts* can be used to compare continuous values across two or more categorical ones.

- *spineplots* show proportions well, but can be hard to interpret



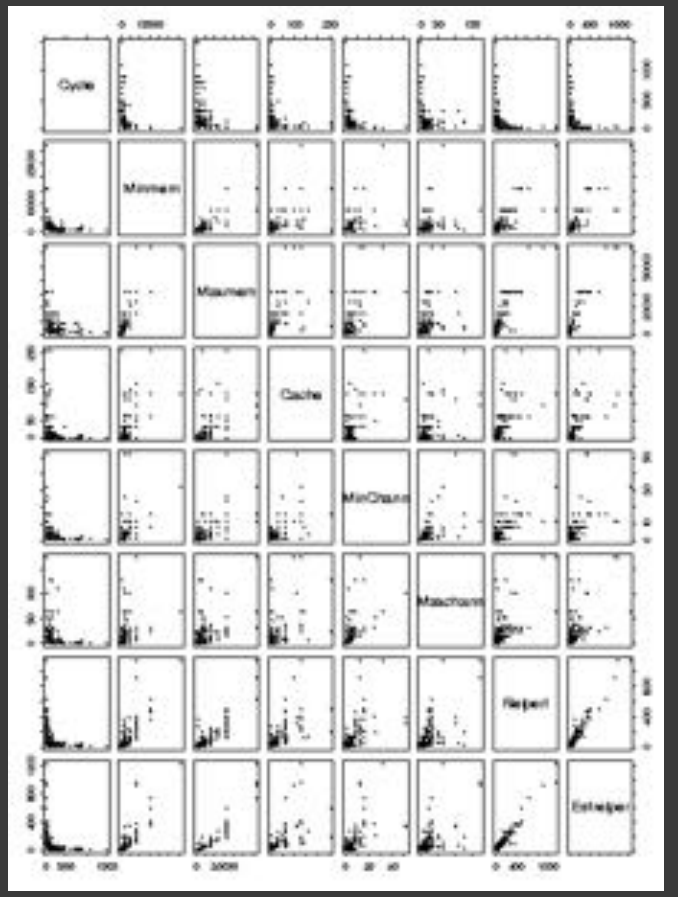


orange=M blue=F

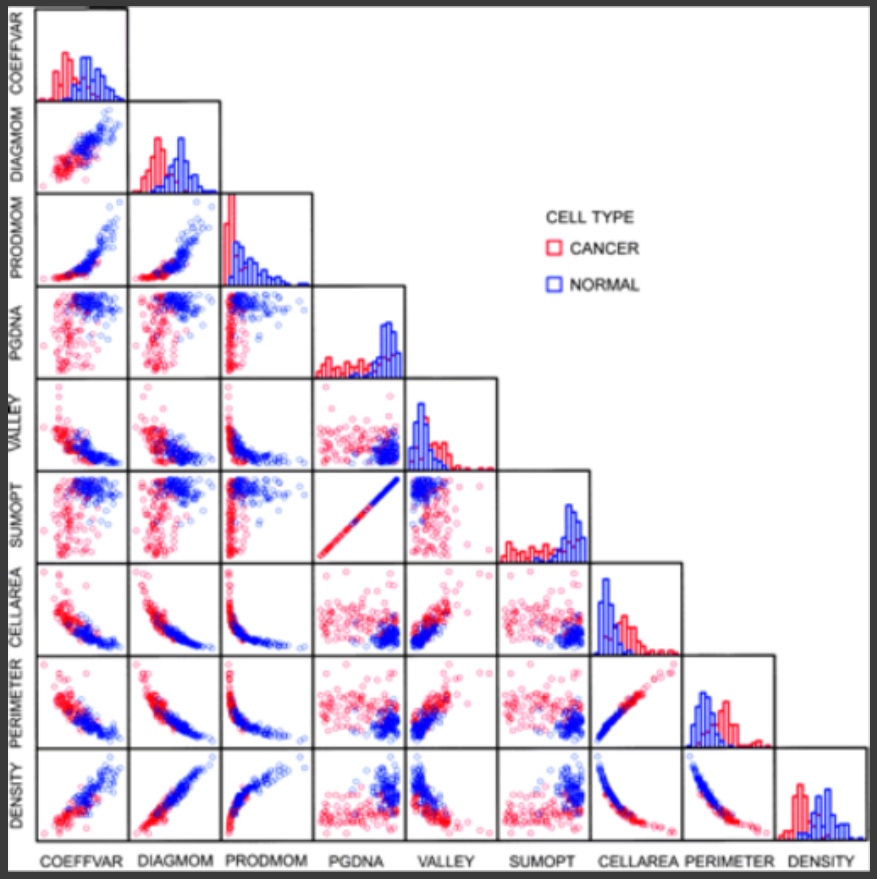


MORE THAN TWO VARIABLES VISUALISATION

- Pairwise scatterplots
- Can be somewhat ineffective for categorical data



MORE THAN TWO VARIABLES VISUALISATION

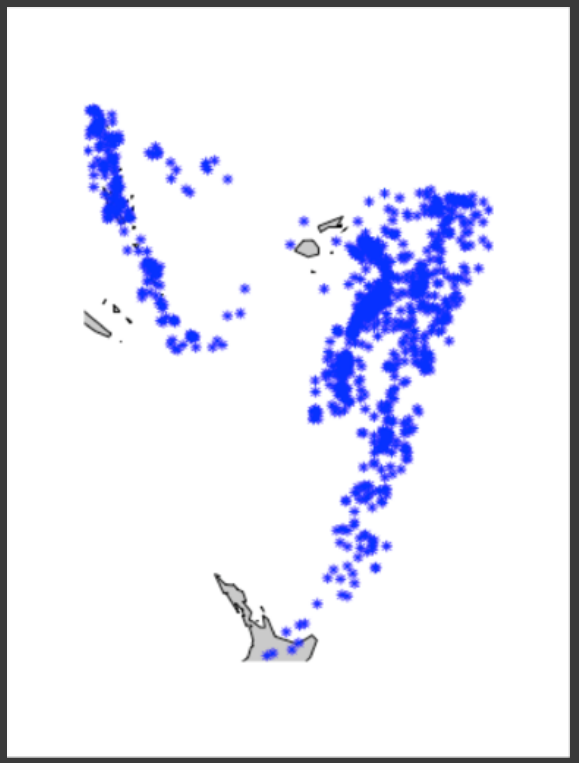


MULTIVARIATE: MORE THAN TWO VARIABLES

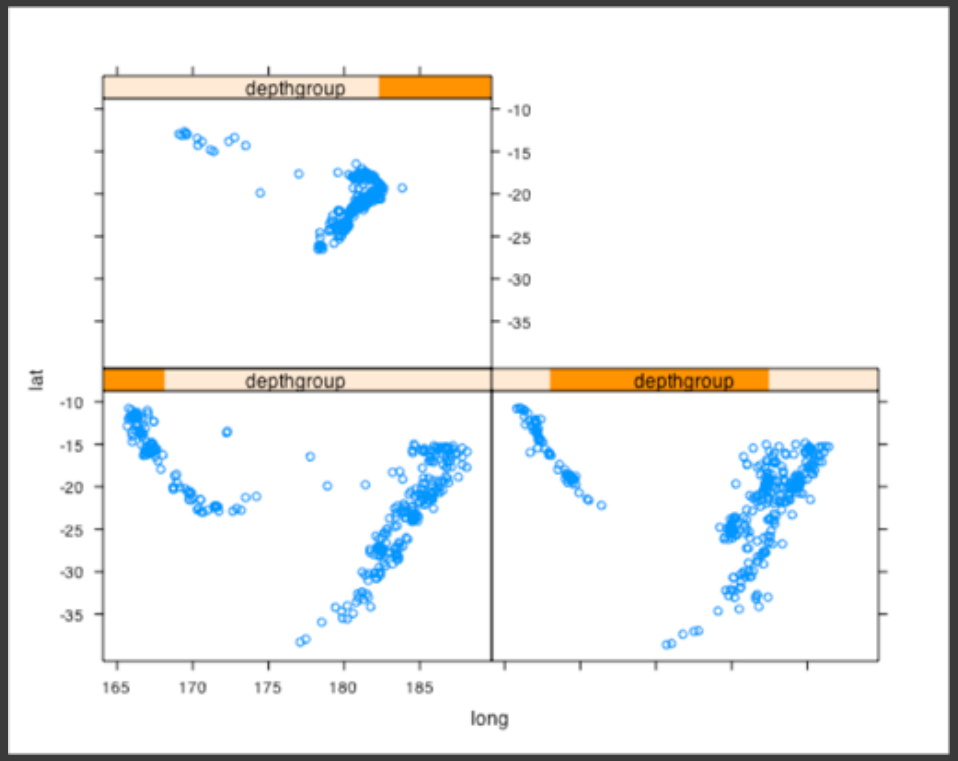
Get creative!

- Conditioning on variables
  - trellis or lattice plots
  - Cleveland models on human perception, all based on conditioning
  - Infinite possibilities

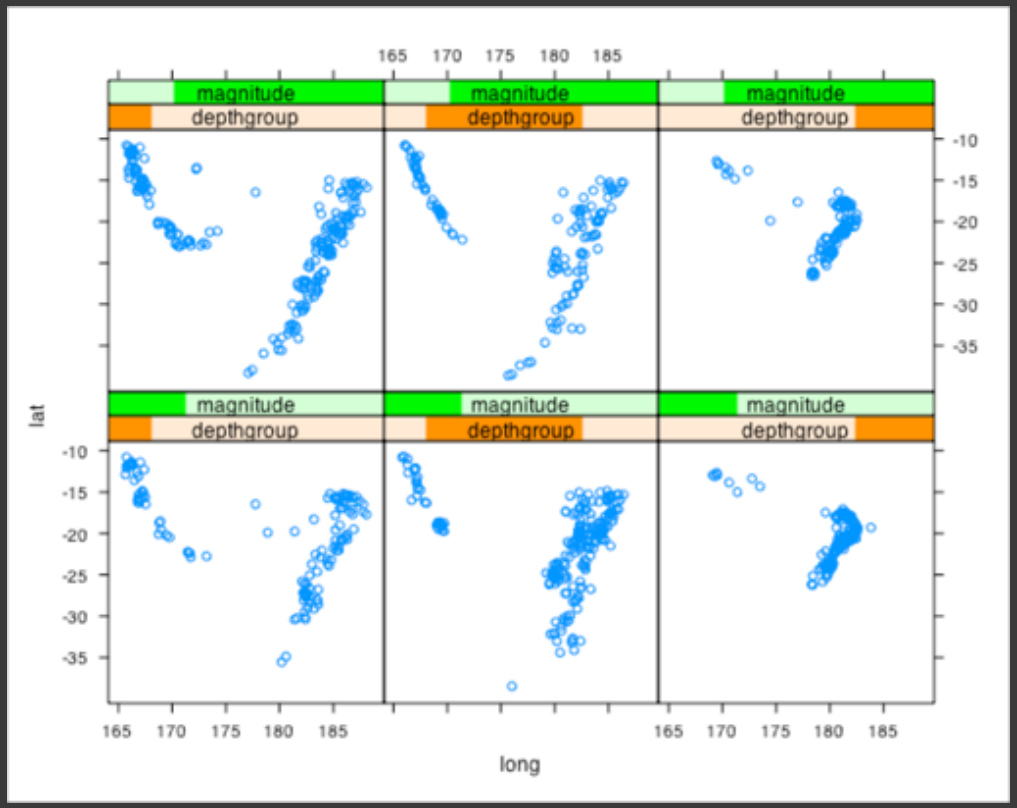
- Earthquake data:
  - locations of 1000 seismic events of MB > 4.0. The events occurred in a cube near Fiji since 1964
  - Data collected on the severity of the earthquake



MULTIVARIATE: MORE THAN TWO VARIABLES



MULTIVARIATE: MORE THAN TWO VARIABLES

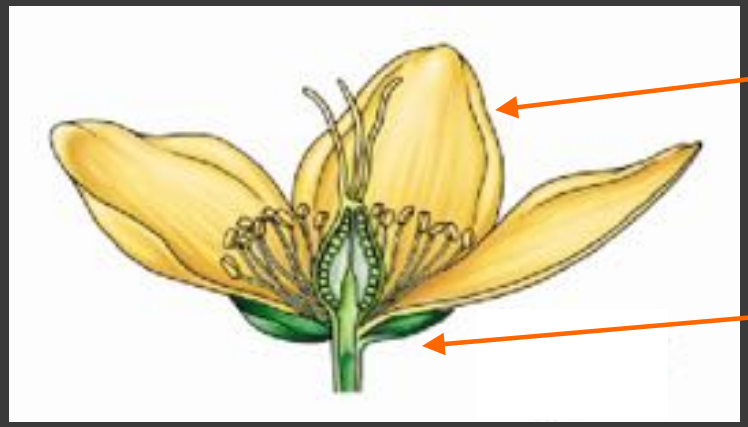


MULTIVARIATE: MORE THAN TWO VARIABLES

How many dimensions are represented here?

Andrew Gelman blog 7/15/2009

MULTIVARIATE VIS: PARALLEL COORDINATES

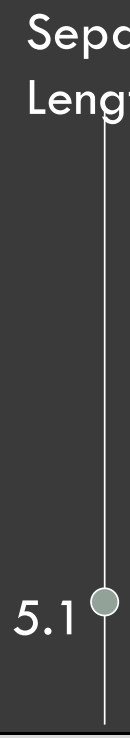


Petal, a non-reproductive part of the flower

Sepal, a non-reproductive part of the flower

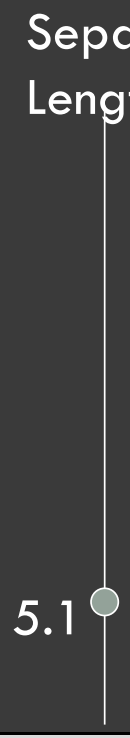
The famous iris data!

PARALLEL COORDINATES



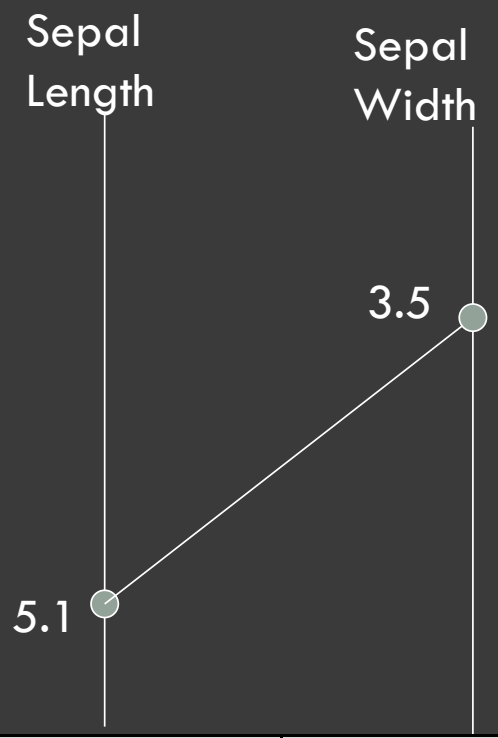
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

PARALLEL COORDINATES



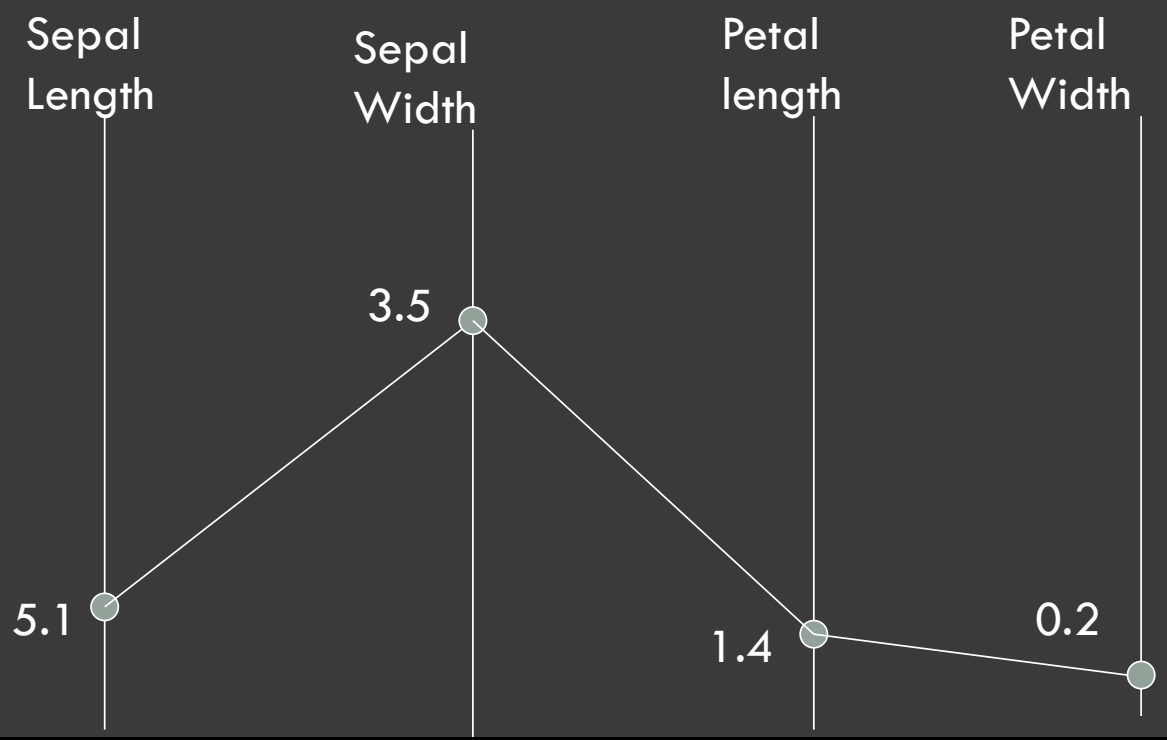
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

PARALLEL COORDINATES: 2D



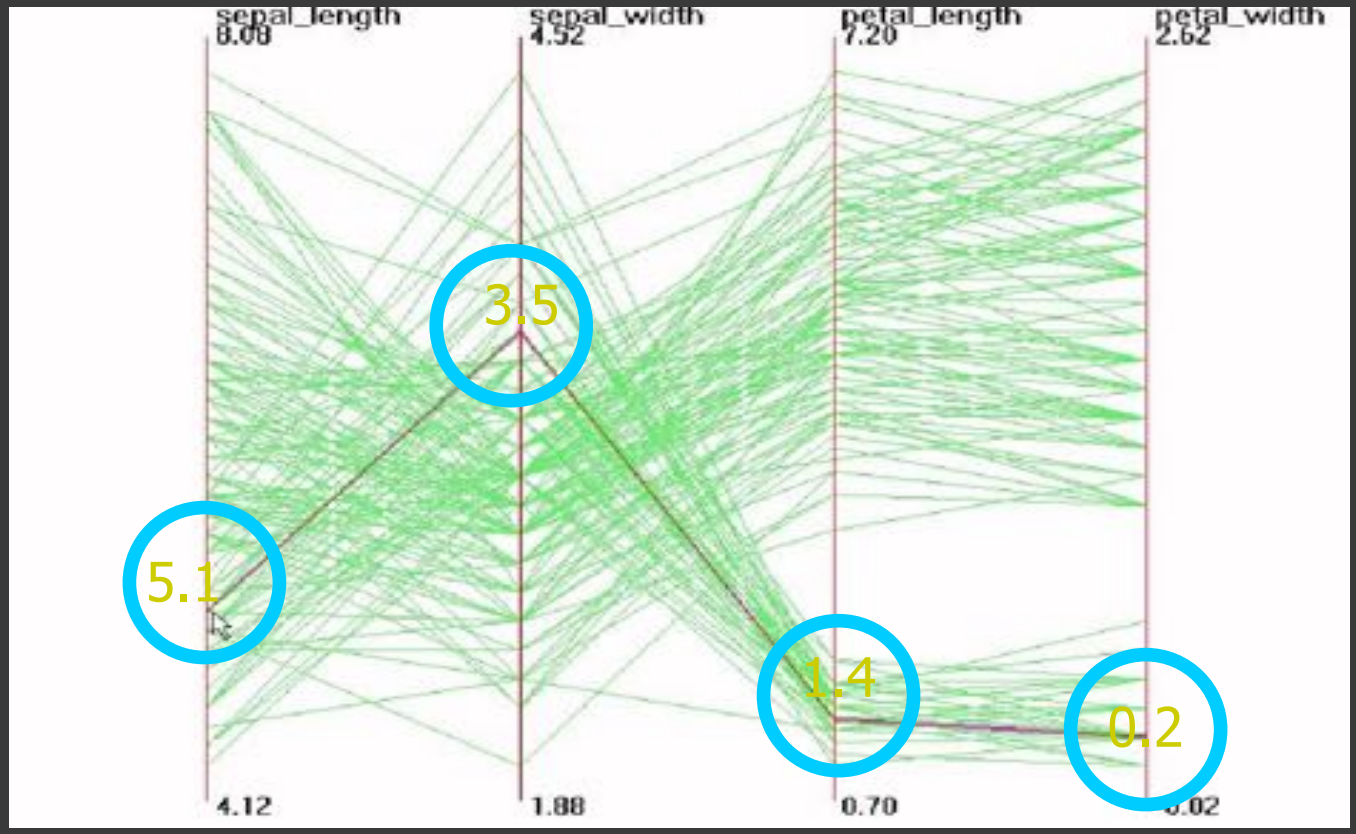
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

PARALLEL COORDINATES: 4 D

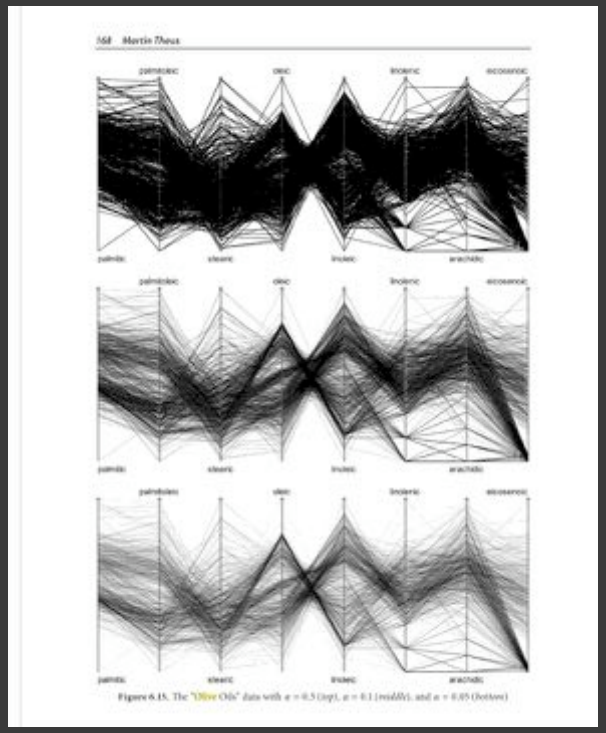


sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

PARALLEL VISUALIZATION OF IRIS DATA



MULTIVARIATE: PARALLEL COORDINATES

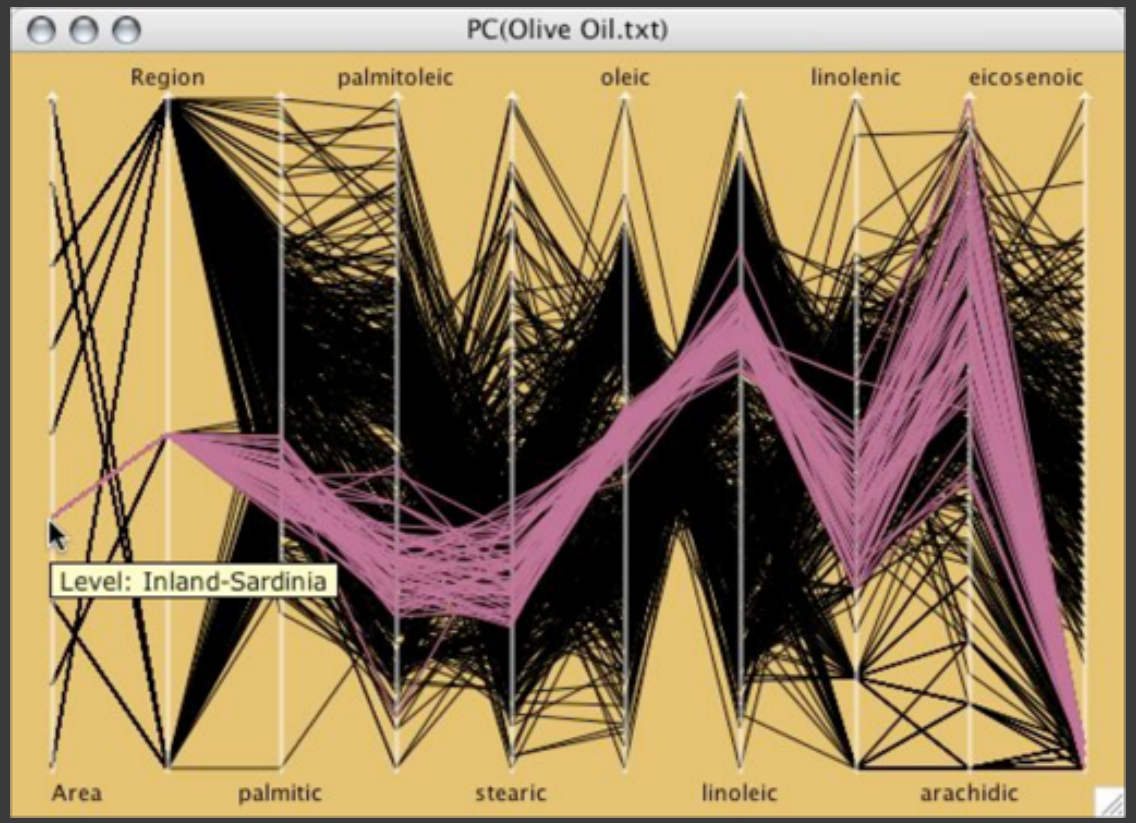


Alpha blending can be effective

Courtesy Unwin, Theus, Hofmann

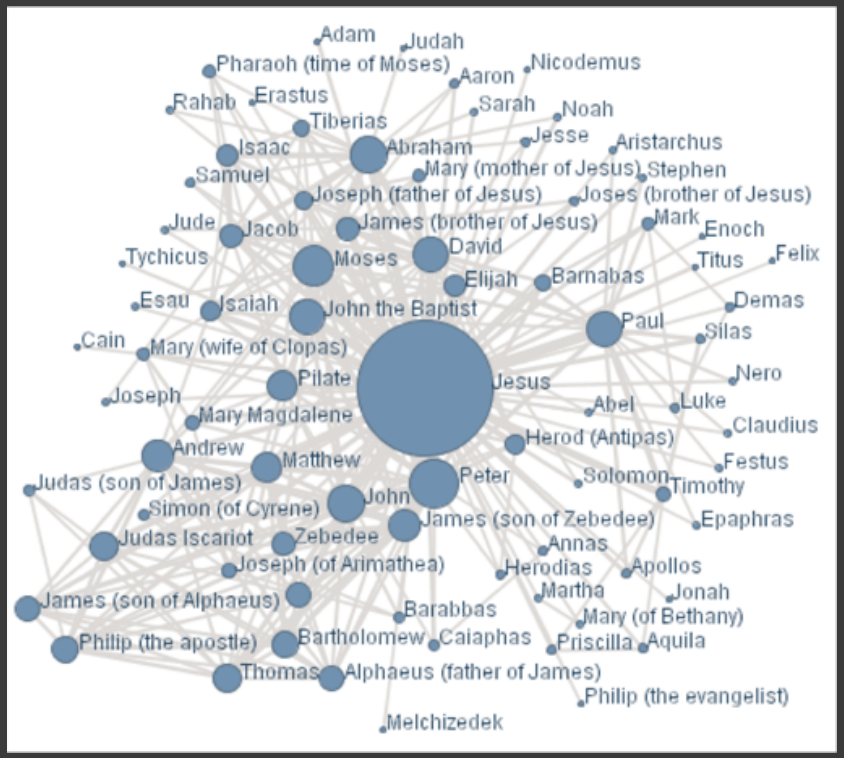
PARALLEL COORDINATES

Useful in an interactive setting



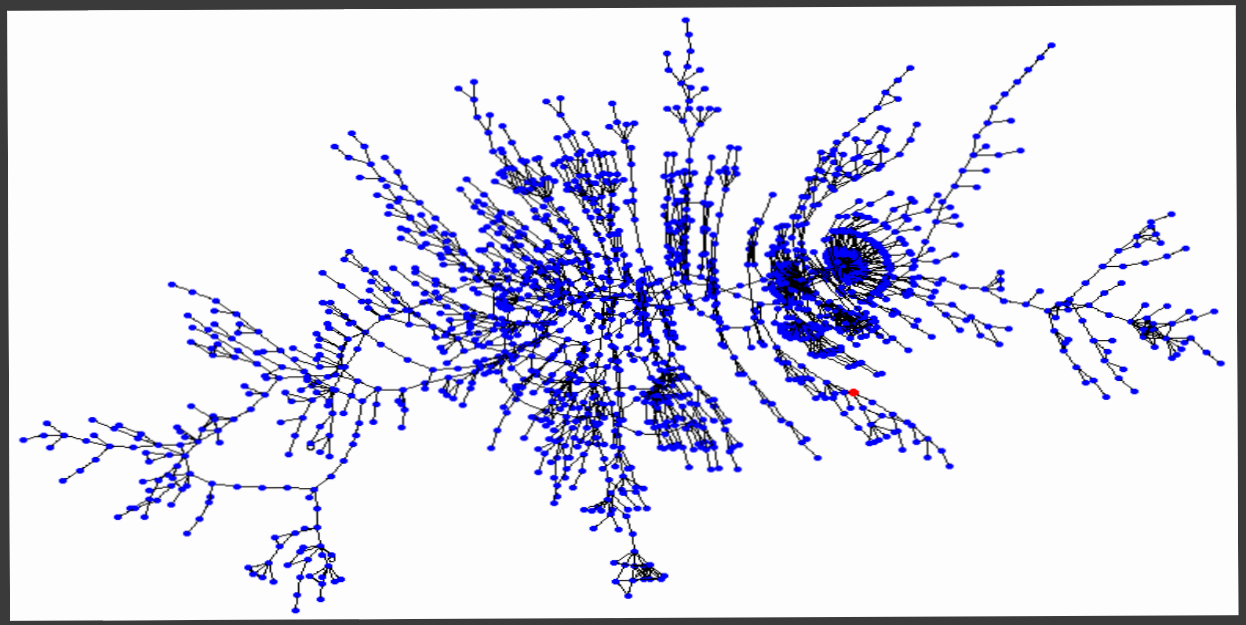
NETWORKS AND GRAPHS

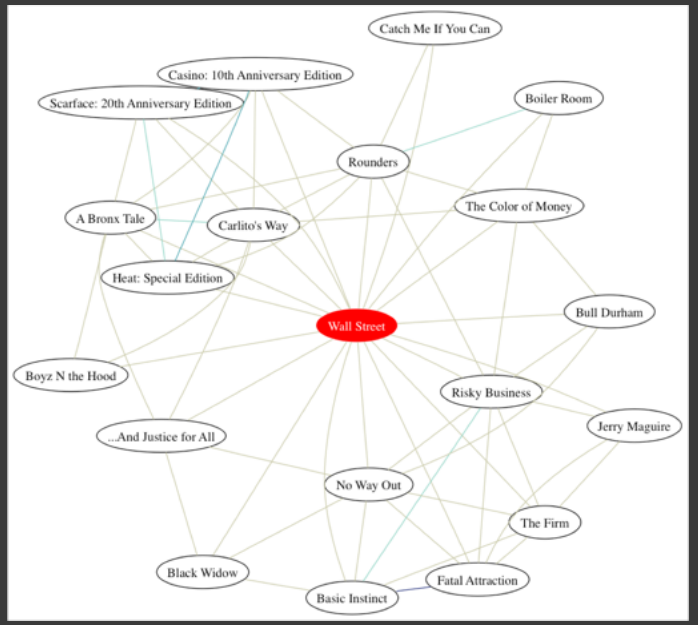
Visualizing networks is helpful, even if is not obvious that a network exists



NETWORK VISUALIZATION

Graphviz (open source software) is a nice layout tool for big and small graphs

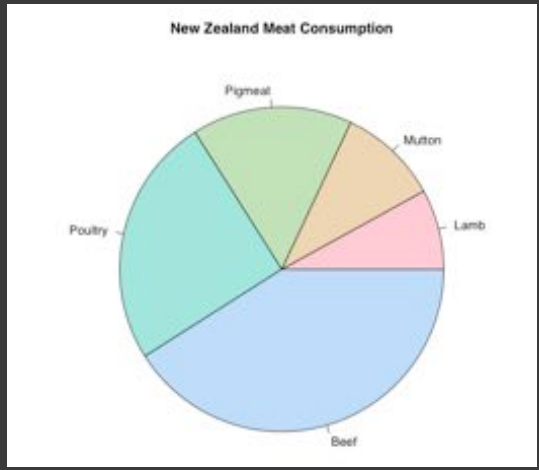




OTHER VISUALISATION TECHNIQUES

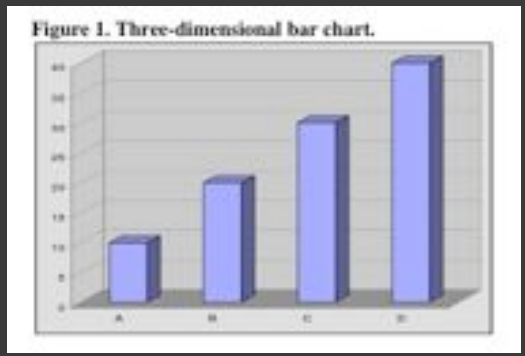
pie charts

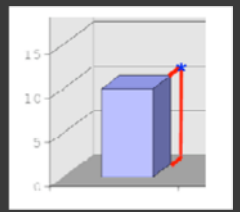
- very popular
- good for showing simple relations of proportions
- Human perception not good at comparing arcs
- barplots, histograms usually better (but less pretty)



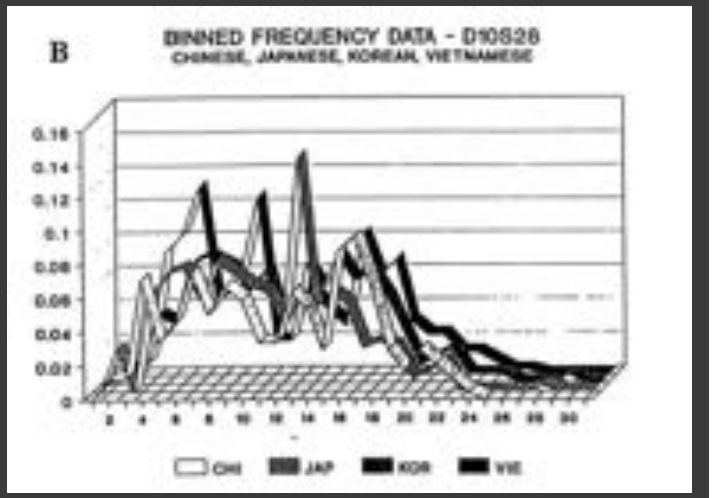
3D

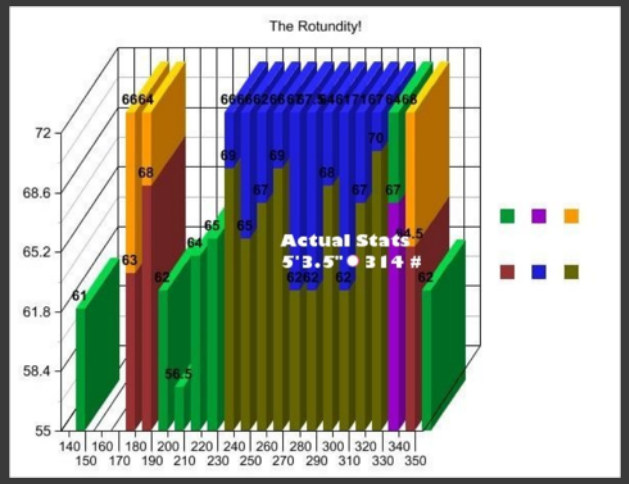
- nice to be able to show three dimensions
- hard to do well
- often done poorly
- 3d best shown through “spinning” in 2D
  - uses various types of projecting into 2D
  - http://www.stat.tamu.edu/~west/bradley/

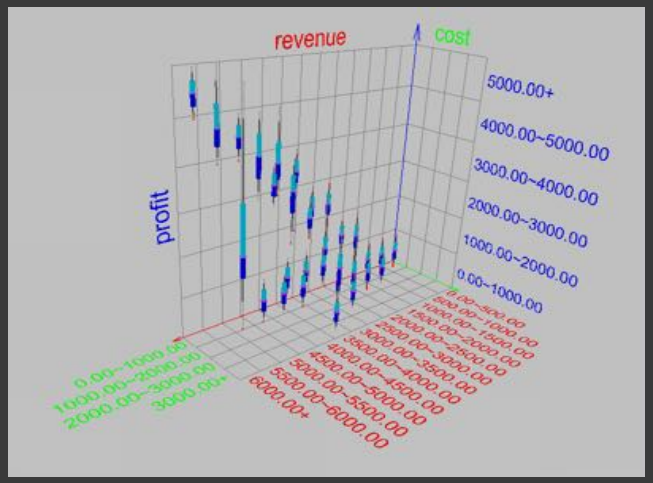


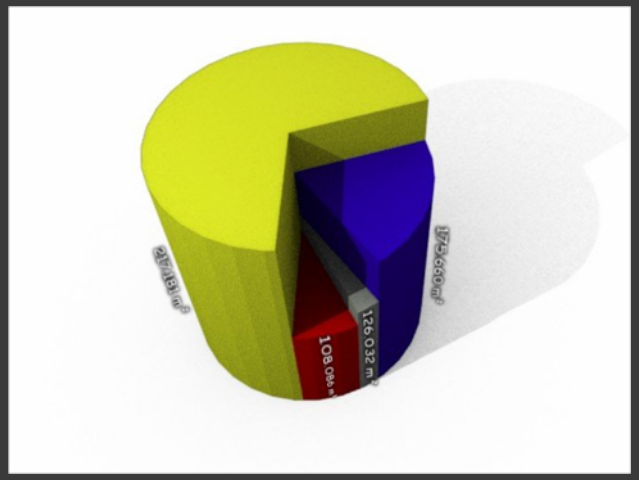


OTHER VISUALISATION TECHNIQUES









DIMENSION REDUCTION

- One way to visualize high dimensional data is to reduce it to 2 or 3 dimensions
  - Variable selection
    - e.g. stepwise
  - Principle Components
    - find linear projection onto p-space with maximal variance
  - Multi-dimensional scaling
    - takes a matrix of (dis)similarities and embeds the points in p-dimensional space to retain those similarities