Exploratory Data Analysis and Data Visualization

Assignment: Read Chapter 2 in Doing Data Science

-> Read the code examples discussed in the book

Outline

- Data Science Cycle
- EDA
- Intro to Data Storage

Data Science Cycle

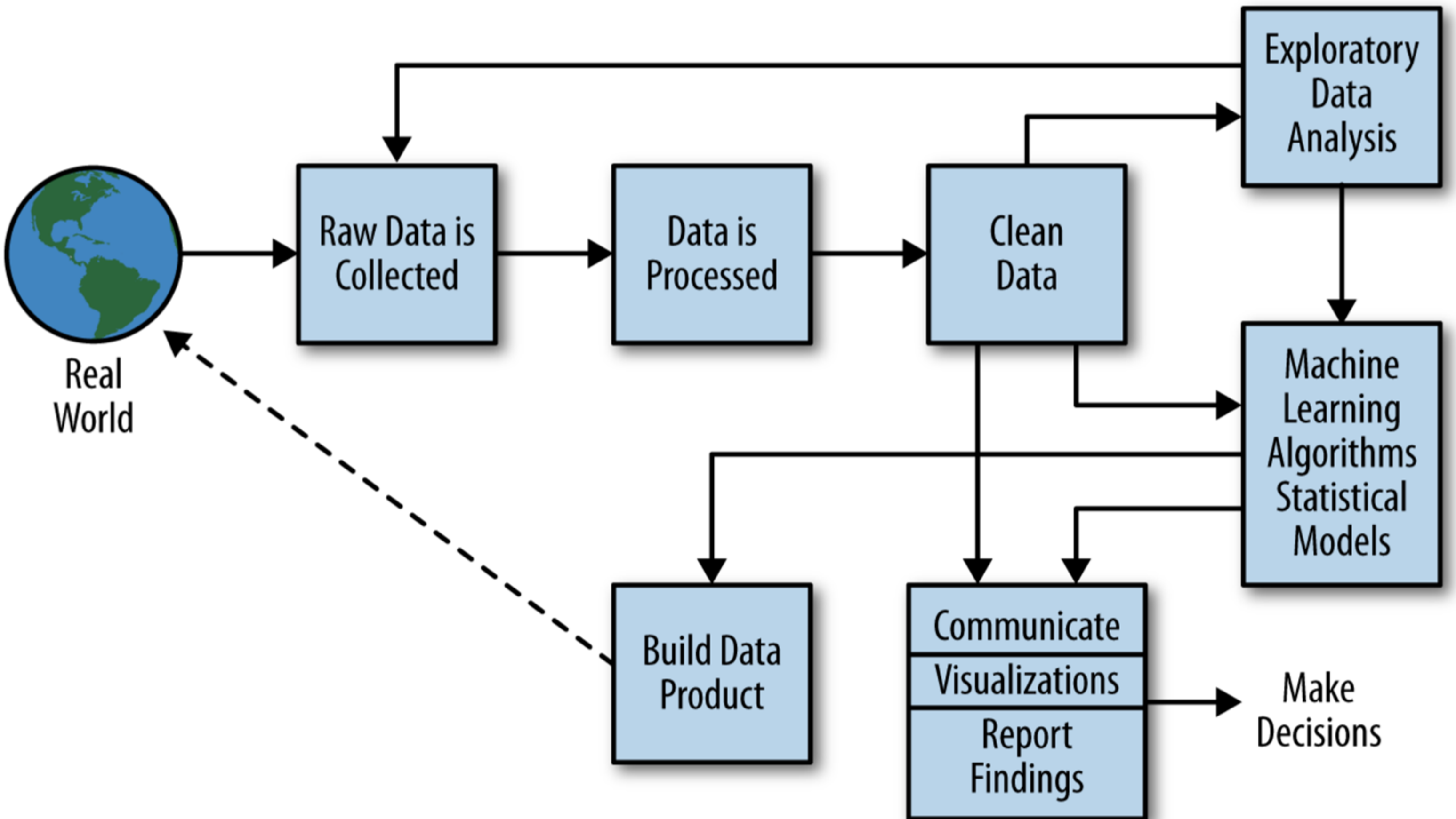


Figure from book

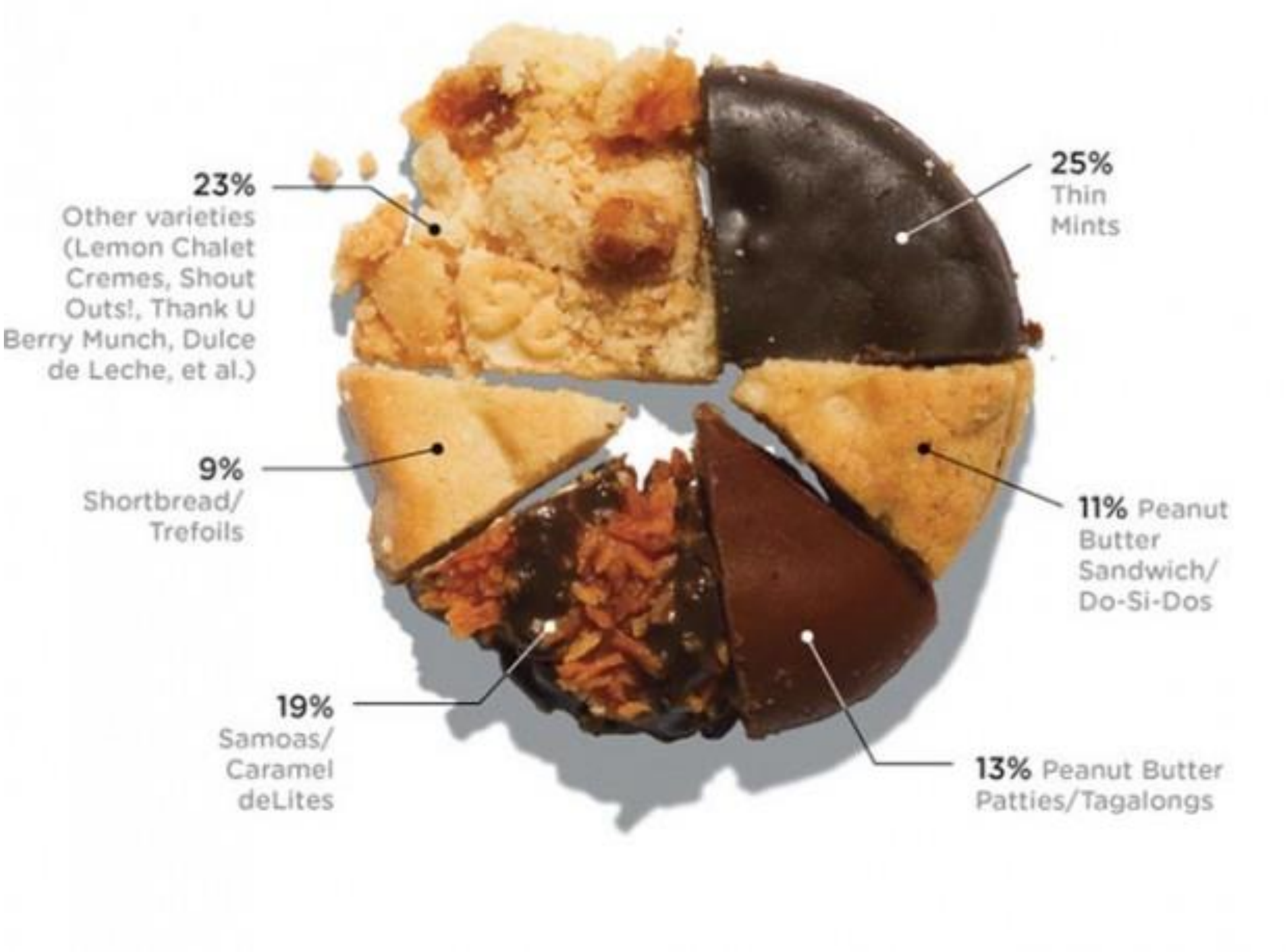
EDA and Visualization

- Exploratory Data Analysis (EDA) and Visualization are very important steps in any analysis task.

- get to know your data!
  - distributions (symmetric, normal, skewed)
  - data quality problems
  - outliers
  - correlations and inter-relationships
  - subsets of interest
  - suggest functional relationships

- Sometimes EDA or viz might be the goal!

Data Visualization – cake bakery



Exploratory Data Analysis (EDA)

- Goal: get a general sense of the data
  - means, medians, quantiles, histograms, boxplots
  - You should always look at every variable - you will learn something!

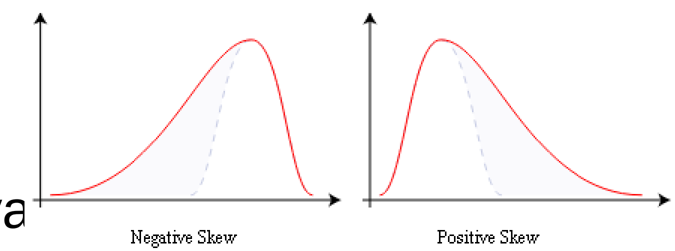
- Think interactive and visual
  - Humans are the best pattern recognizers
  - You can use more than 2 dimensions!
    - x,y,z, space, color, time….

- Especially useful in early stages of data mining
  - detect outliers (e.g. assess data quality)
  - test assumptions (e.g. normal distributions or skewed?)
  - identify useful raw data & transforms (e.g. log(x))

- Bottom line: it is always well worth looking at your data!

Summary Statistics

- *not* visual
- sample statistics of data X
  - mean: $\mu = \sum_i X_i / n$
  - mode: most common value in X
  - median: $\mathbf{X}$=sort(X), median = $\mathbf{X}_{n/2}$ (half below, half above)
  - quartiles of sorted $\mathbf{X}$: Q1 value = $\mathbf{X}_{0.25n}$, Q3 value = $\mathbf{X}_{0.75\,n}$
    - interquartile range: value(Q3) - value(Q1)
    - range: max(X) - min(X) = $\mathbf{X}_n - \mathbf{X}_1$
  - variance: $\sigma^2 = \sum_i (X_i - \mu)^2 / n$
  - skewness: $\sum_i (X_i - \mu)^3 \;/\; [\, (\sum_i (X_i - \mu)^2)^{3/2} \,]$
    - zero if symmetric; right-skewed more common (what kind of data is right skewed?)

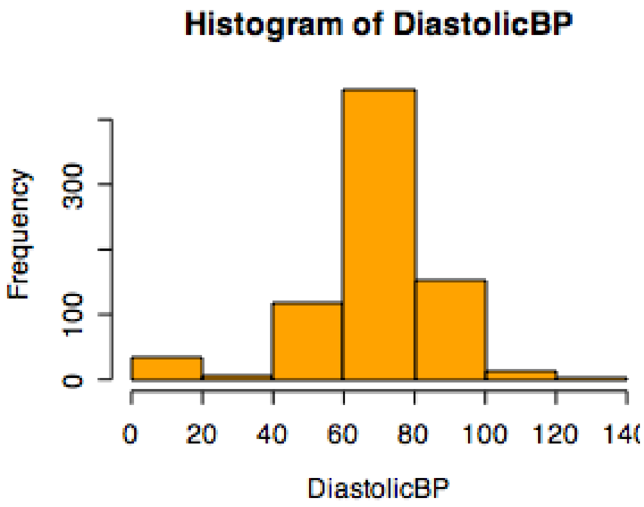


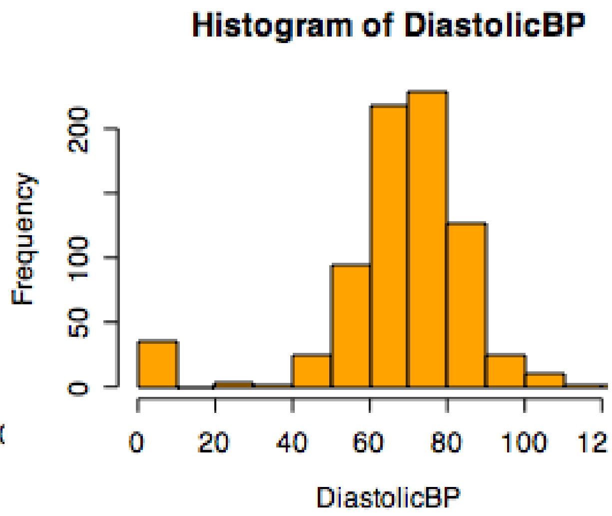
  - number of distinct values for a va
  - Don’t need to report all of thses: Bottom line…do these numbers make sense???

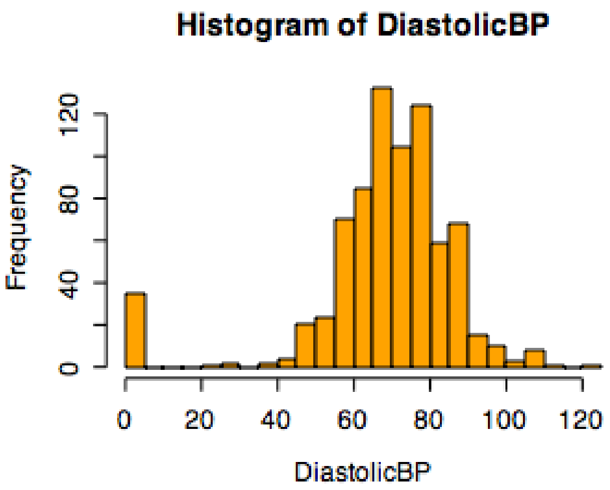
Single Variable Visualization

Histogram:

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







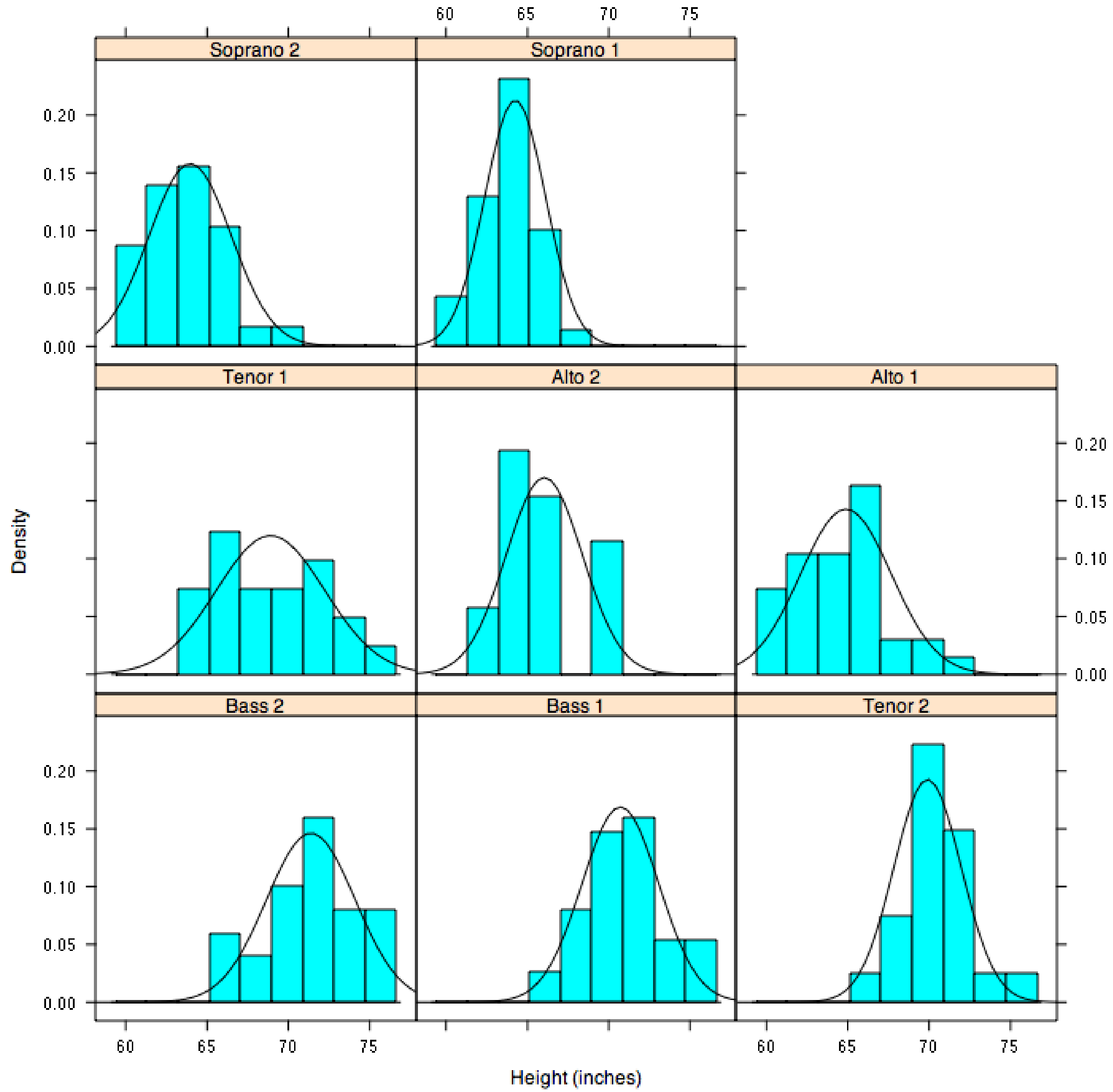
Issues with Histograms

- 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

But be careful with axes and scales!

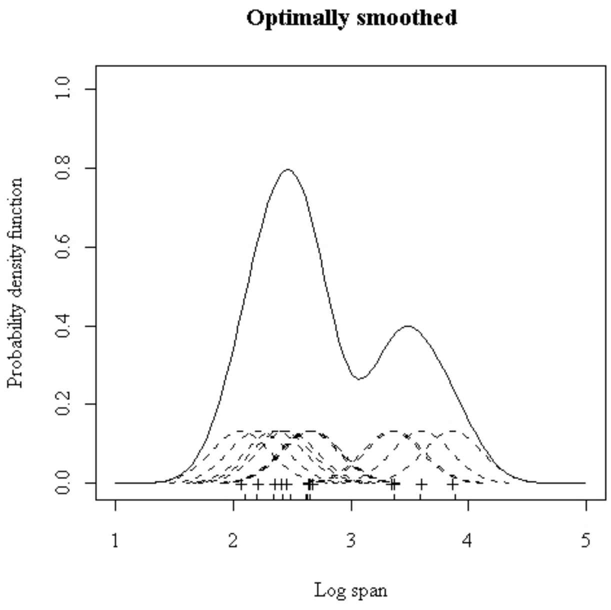


Smoothed Histograms - Density Estimates

- Kernel estimates smooth out the contribution of each datapoint 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

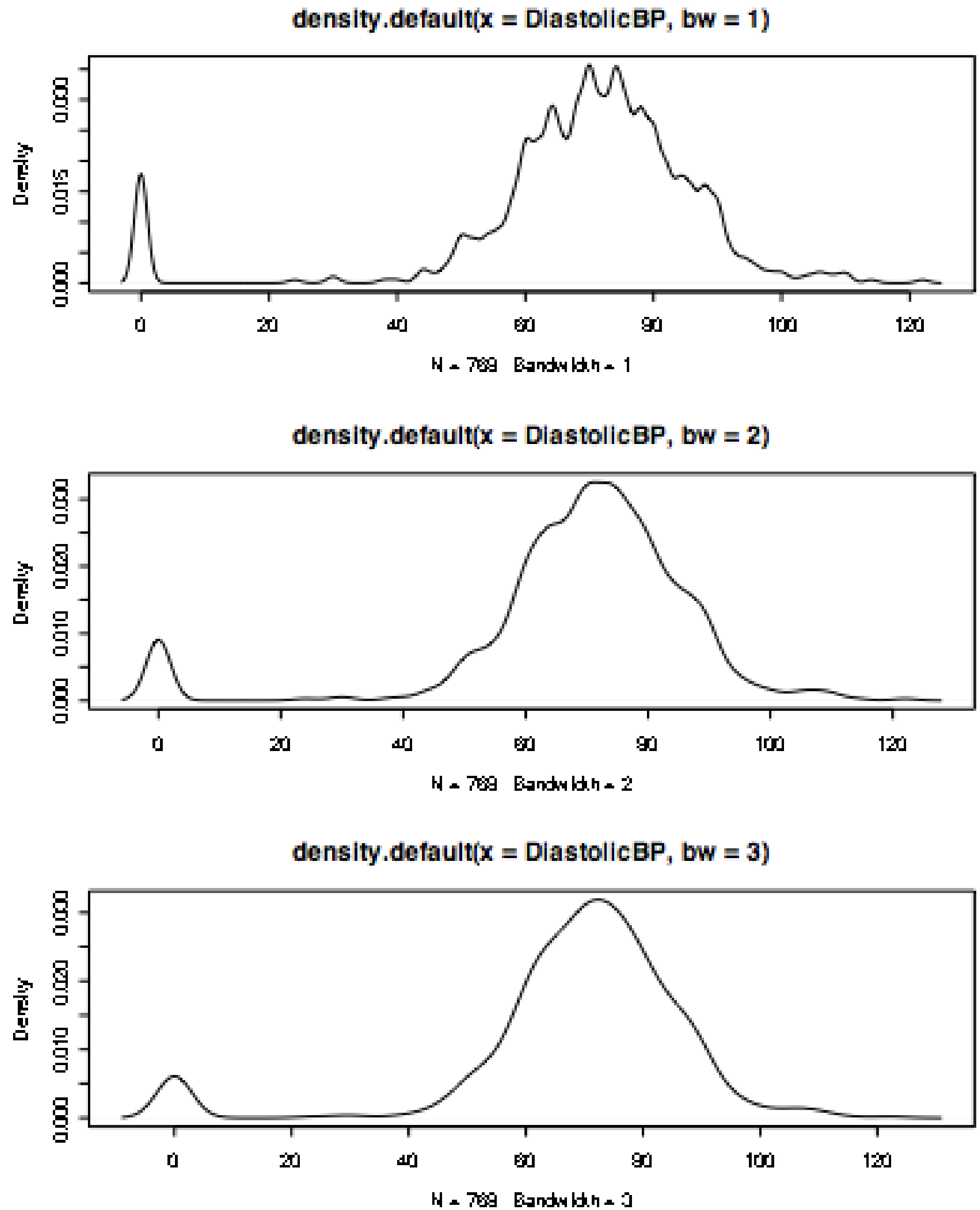


- Gaussian kernel is common:

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

Bandwidth choice is an art

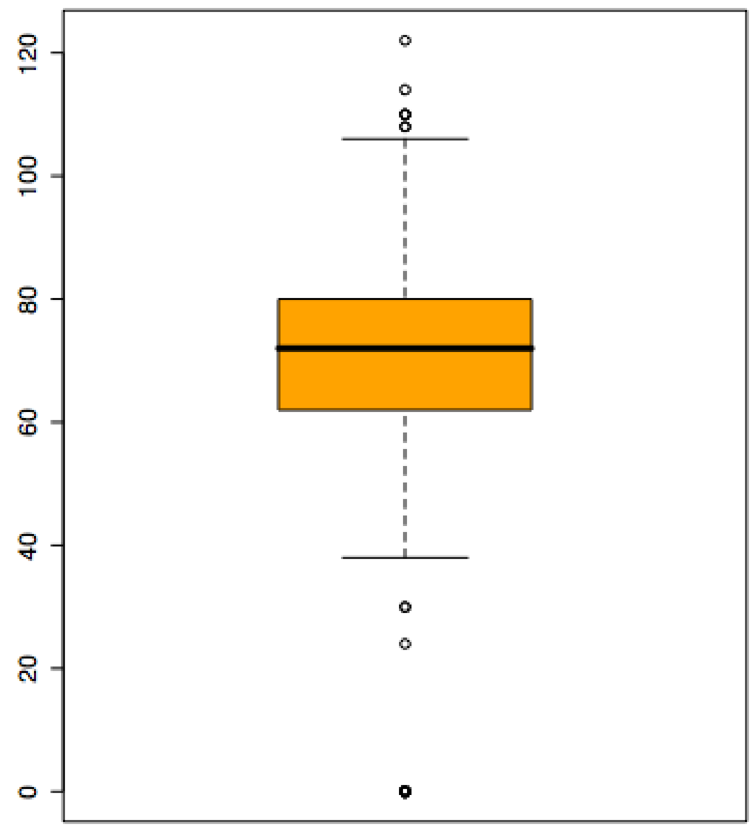
Usually want to try several



Boxplots

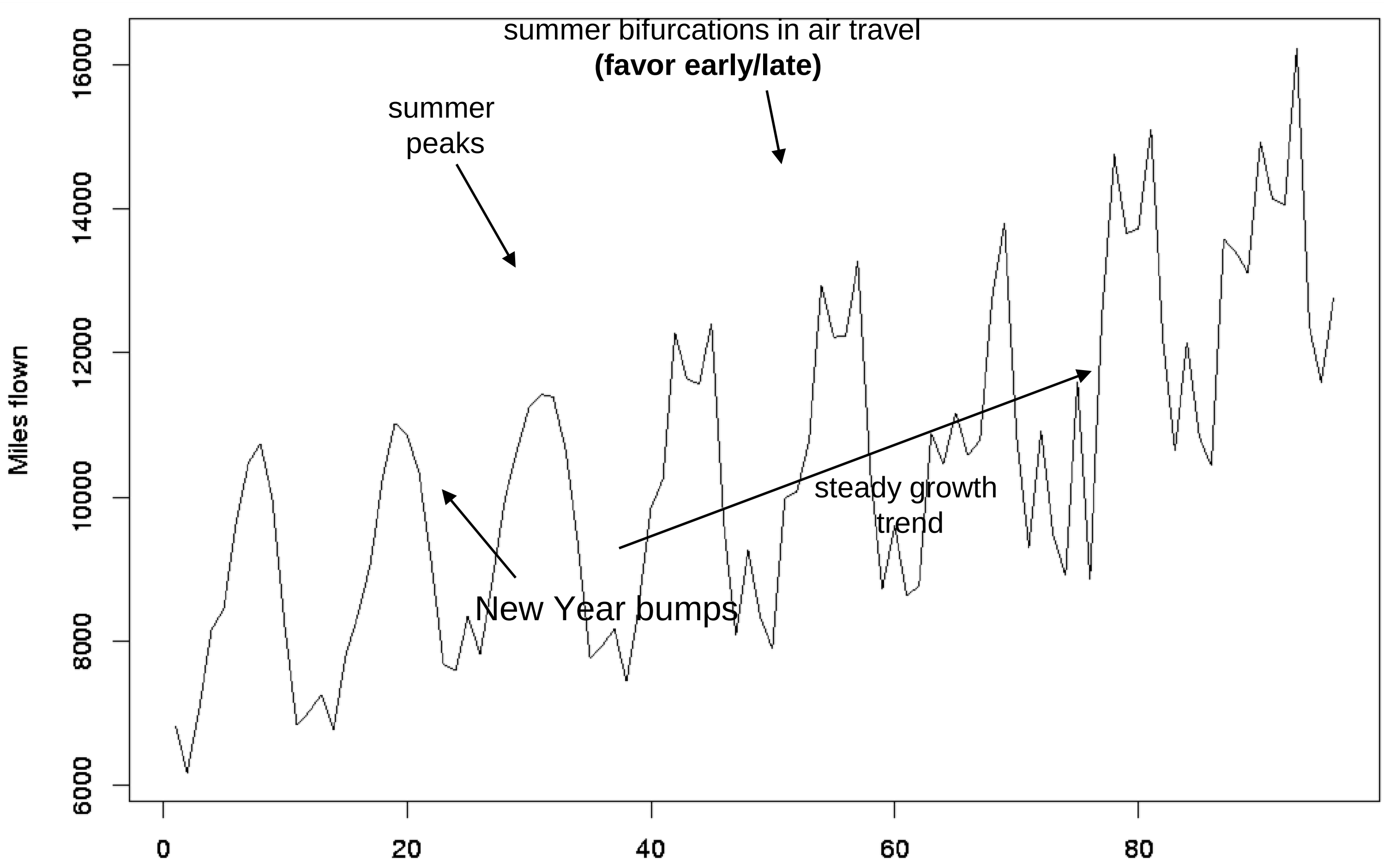
- Shows a lot of information about a variable in one plot
  - Median
  - IQR
  - Outliers
  - Range
  - Skewness

- Negatives
  - Overplotting
  - Hard to tell distributional shape
  - no standard implementation in software (many options for whiskers, outliers)

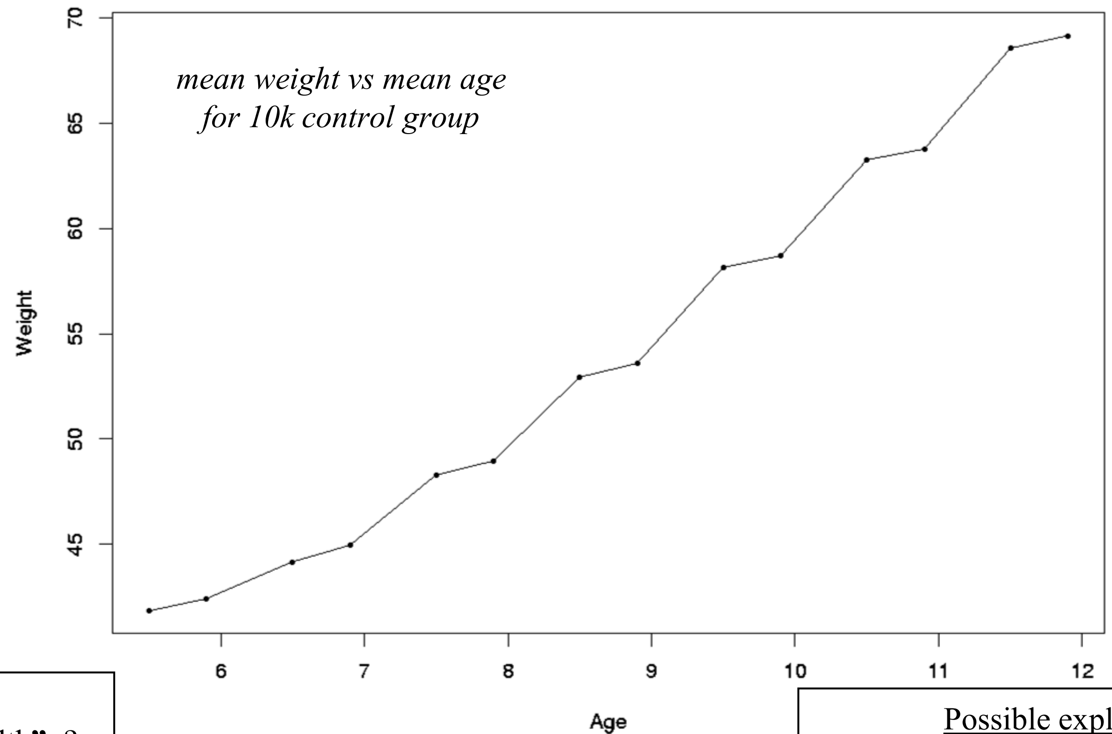


Time Series

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



Time-Series Example 3



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
<u>unexpected pattern</u> (*steps*),
not apparent from raw data table.

<u>Possible explanations</u>:

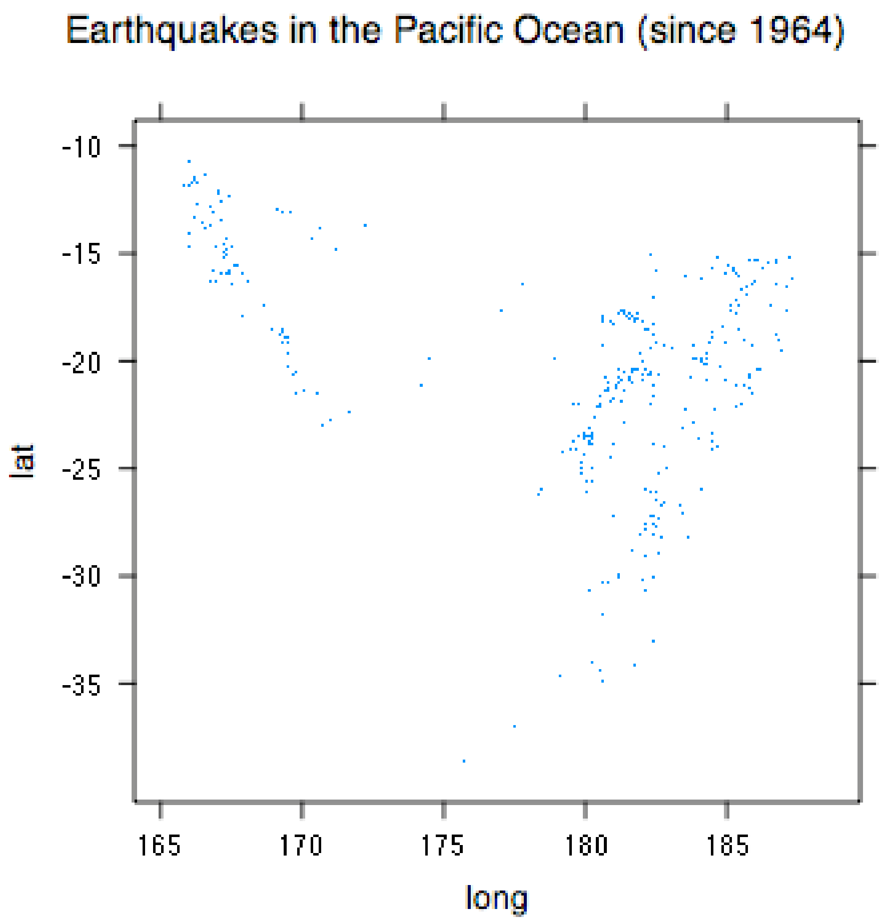
Grow less early in year than later?

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

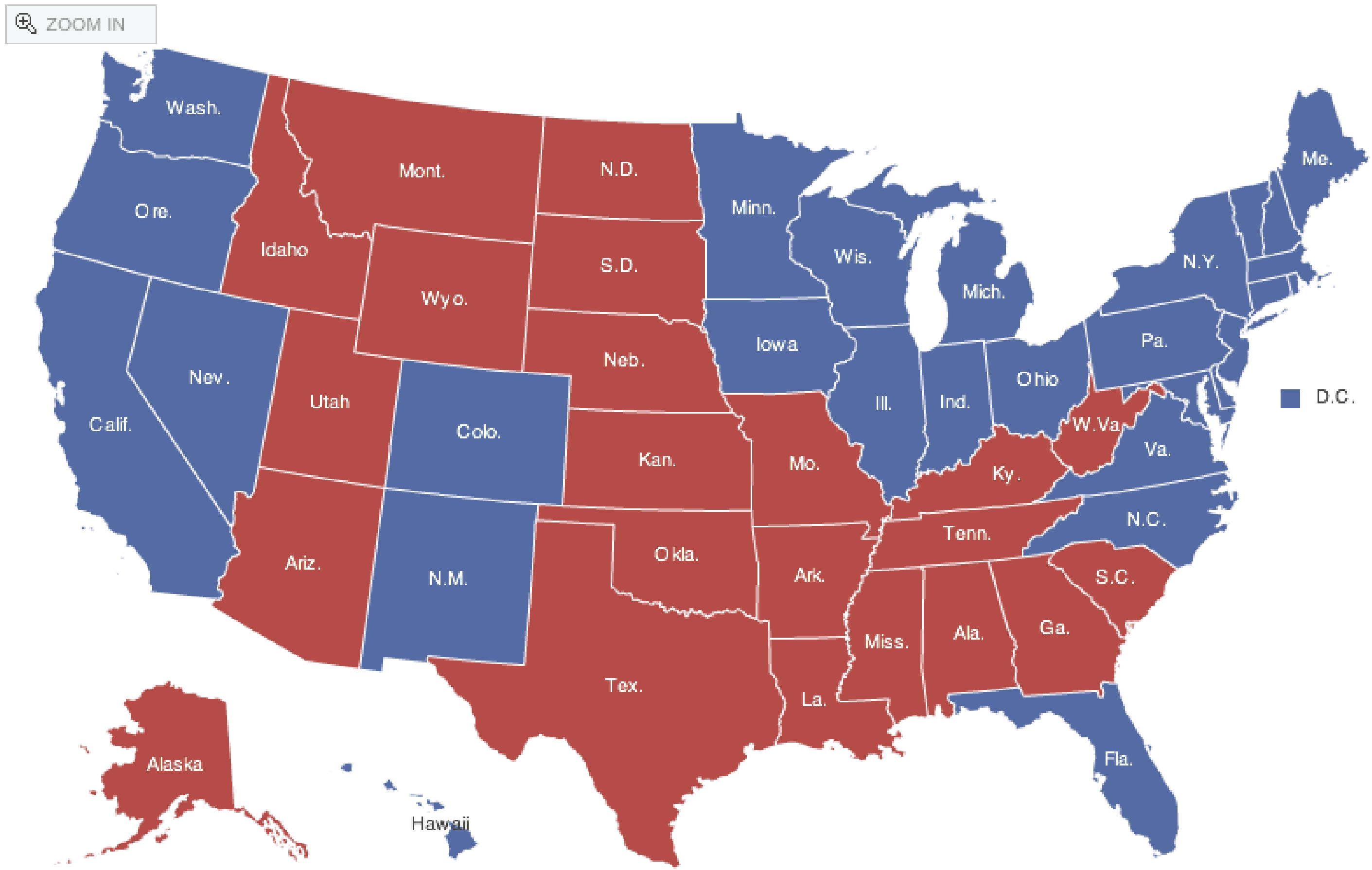
Kids weighed in clothes: summer garb lighter than winter?

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



Spatial data: choropleth Maps

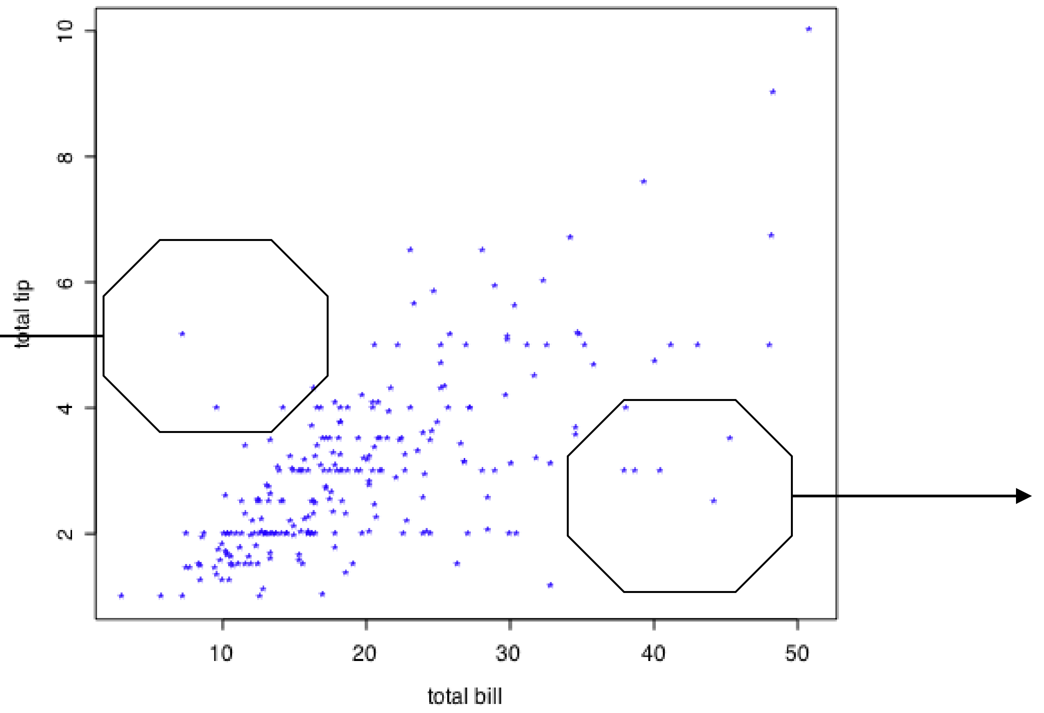


Maps using color shadings to represent numerical values are called chloropleth maps
http://elections.nytimes.com/2008/results/president/map.html

Two Continuous Variables

For two numeric variables, the scatterplot is the obvious choice

interesting?



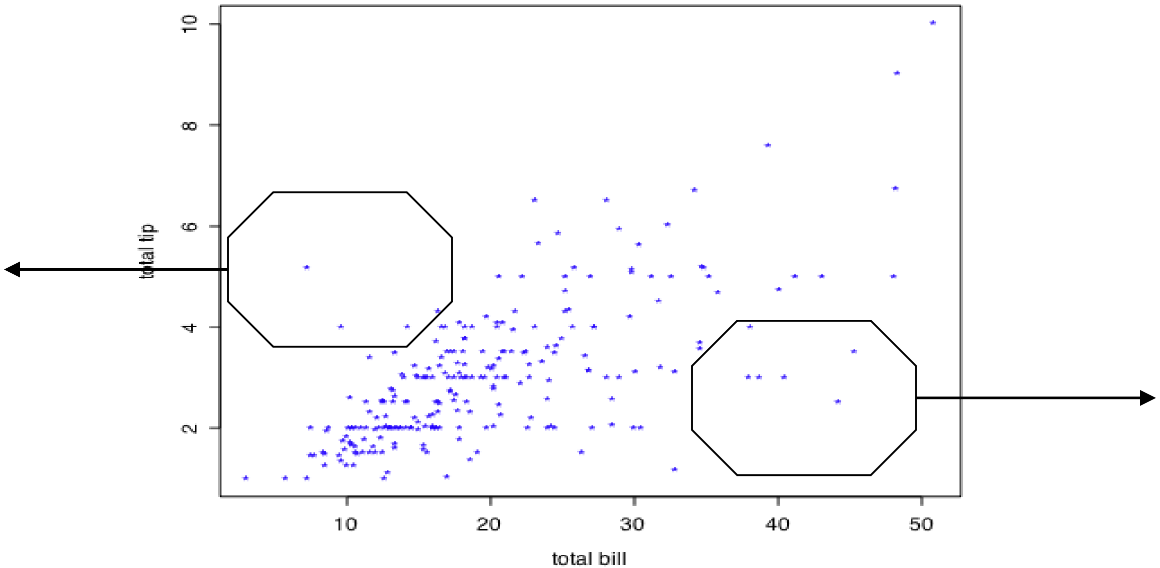
interesting?

2D Scatterplots

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

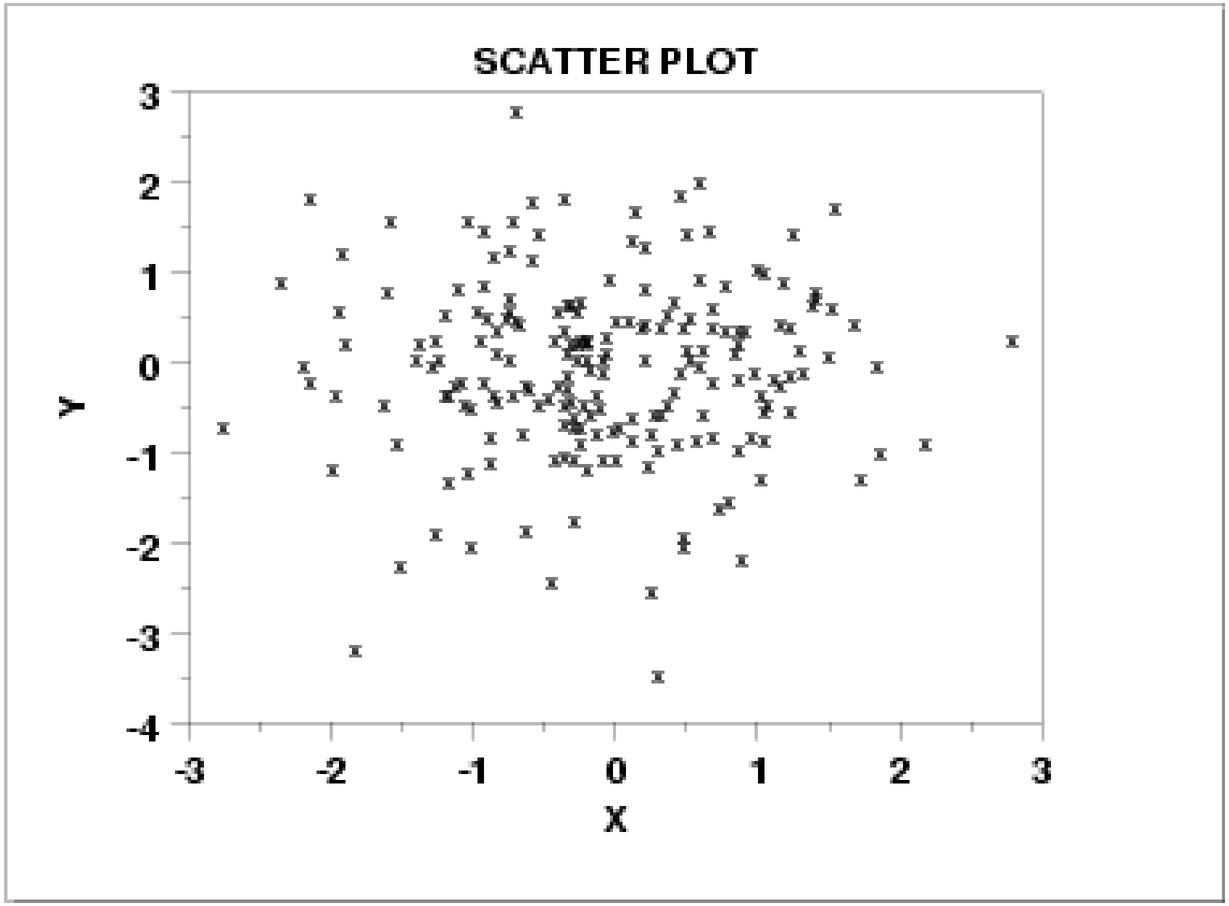
- useful to answer:
  - x,y related?
    - linear
    - quadratic
    - other
  - variance(y) depend on x?
  - outliers present?

interesting?

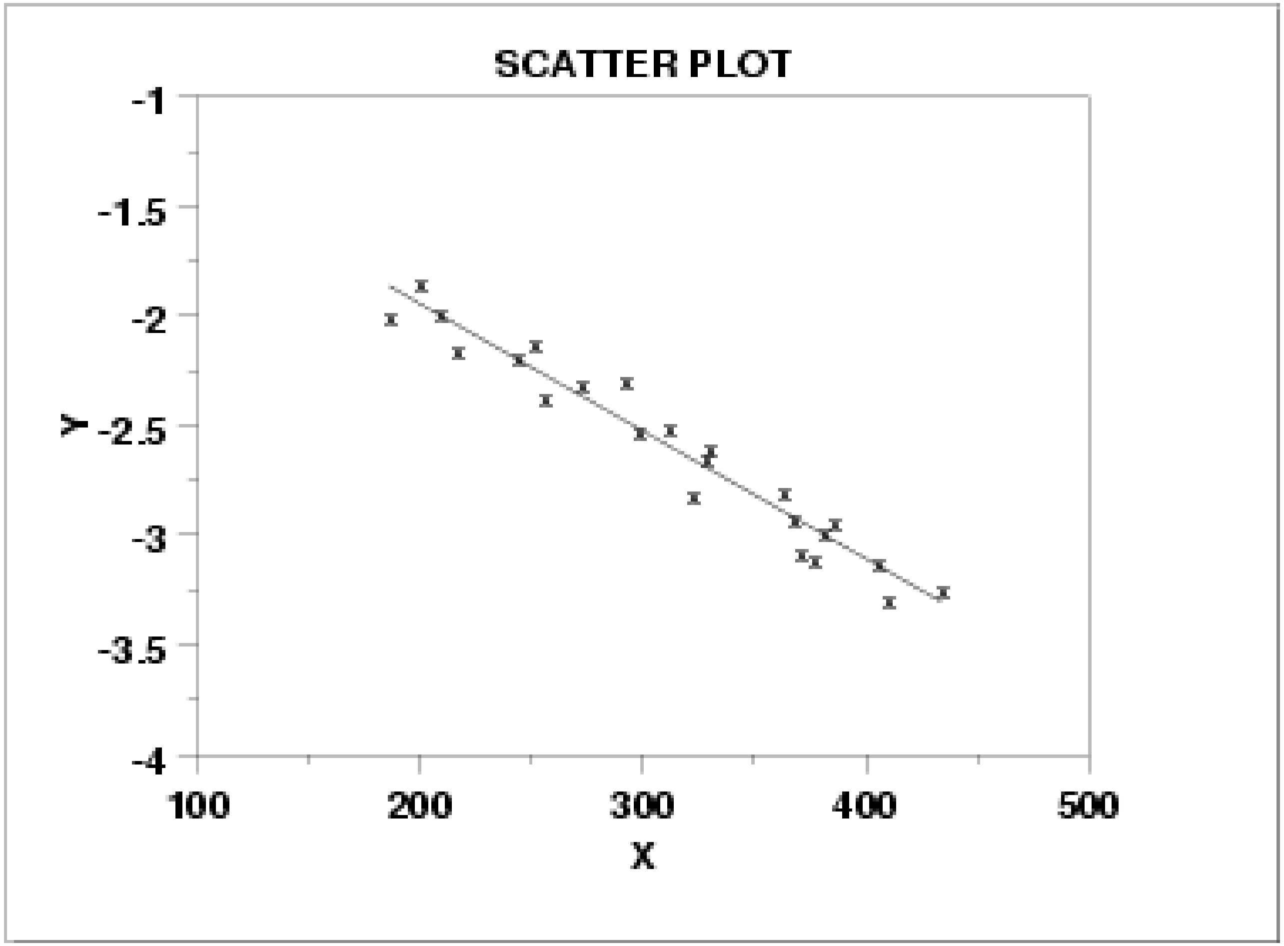


interesting?

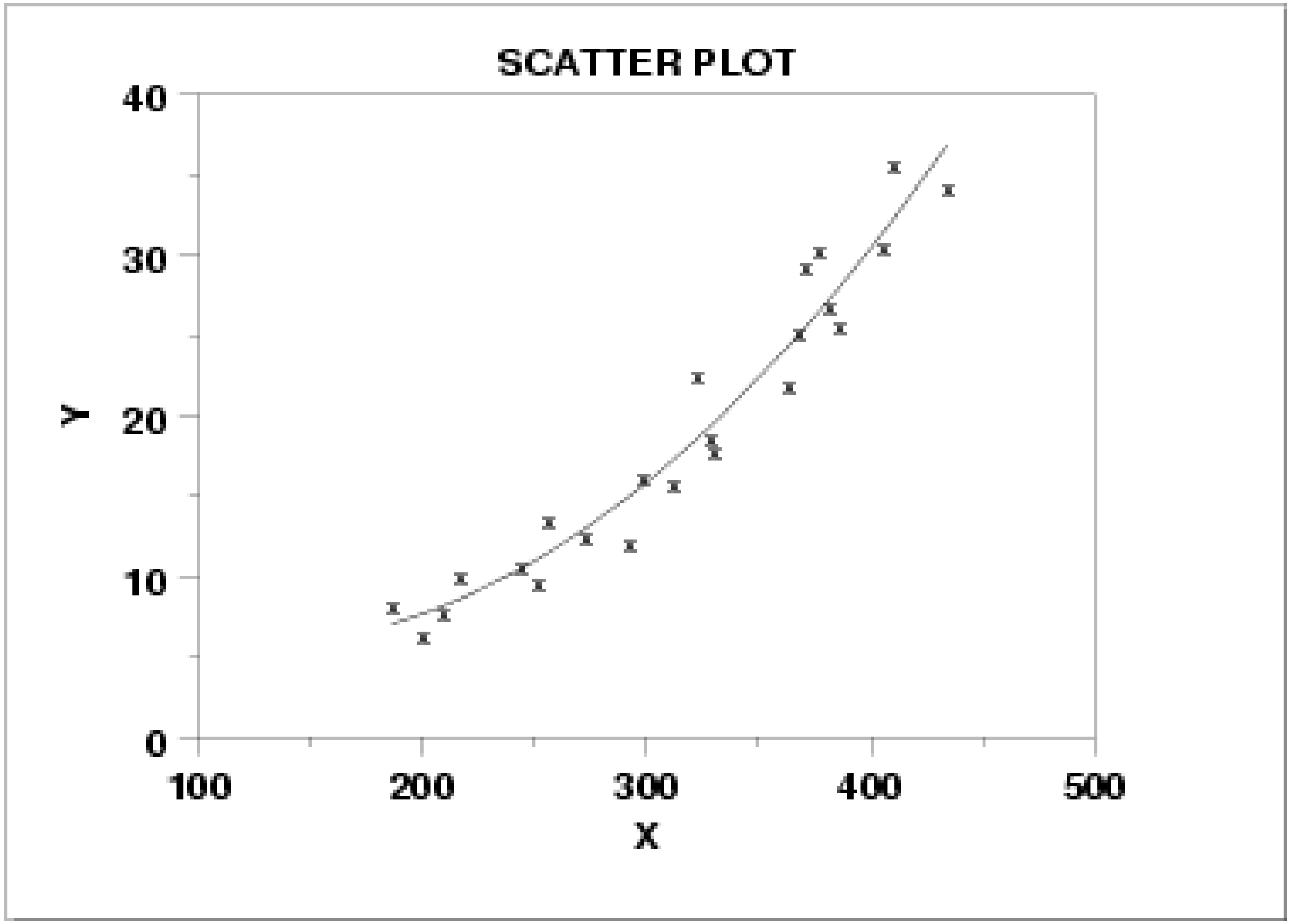
Scatter Plot: No apparent relationship



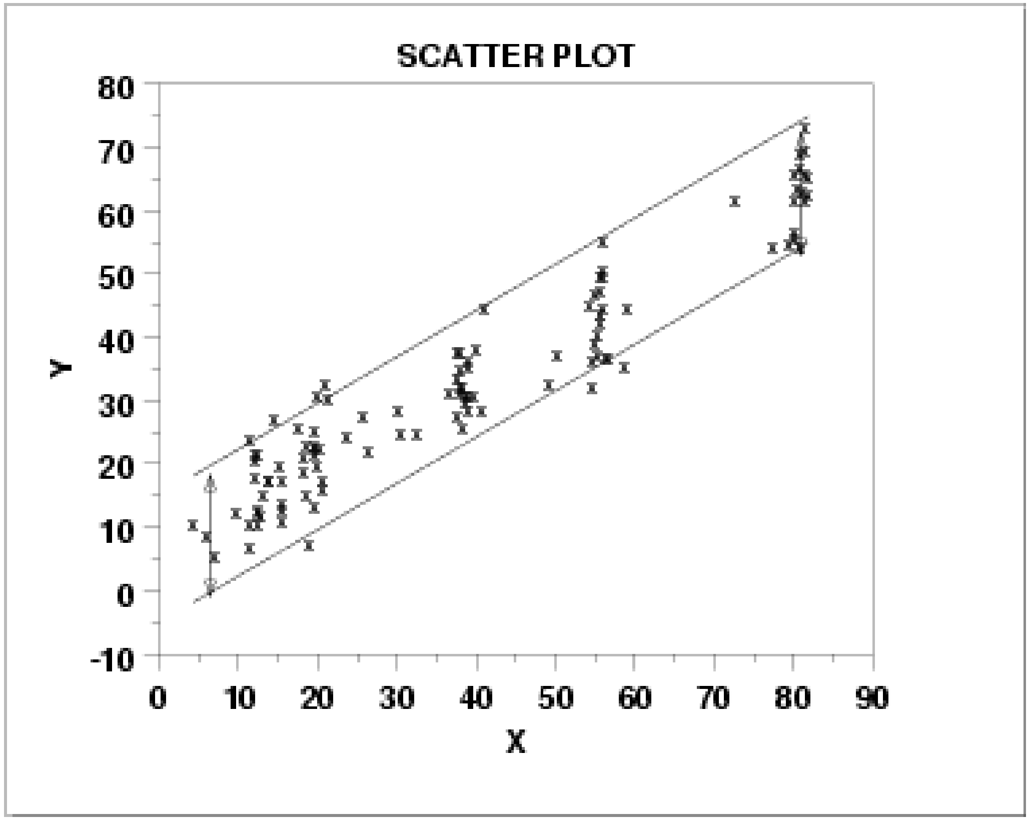
Scatter Plot: Linear relationship



Scatter Plot: Quadratic relationship

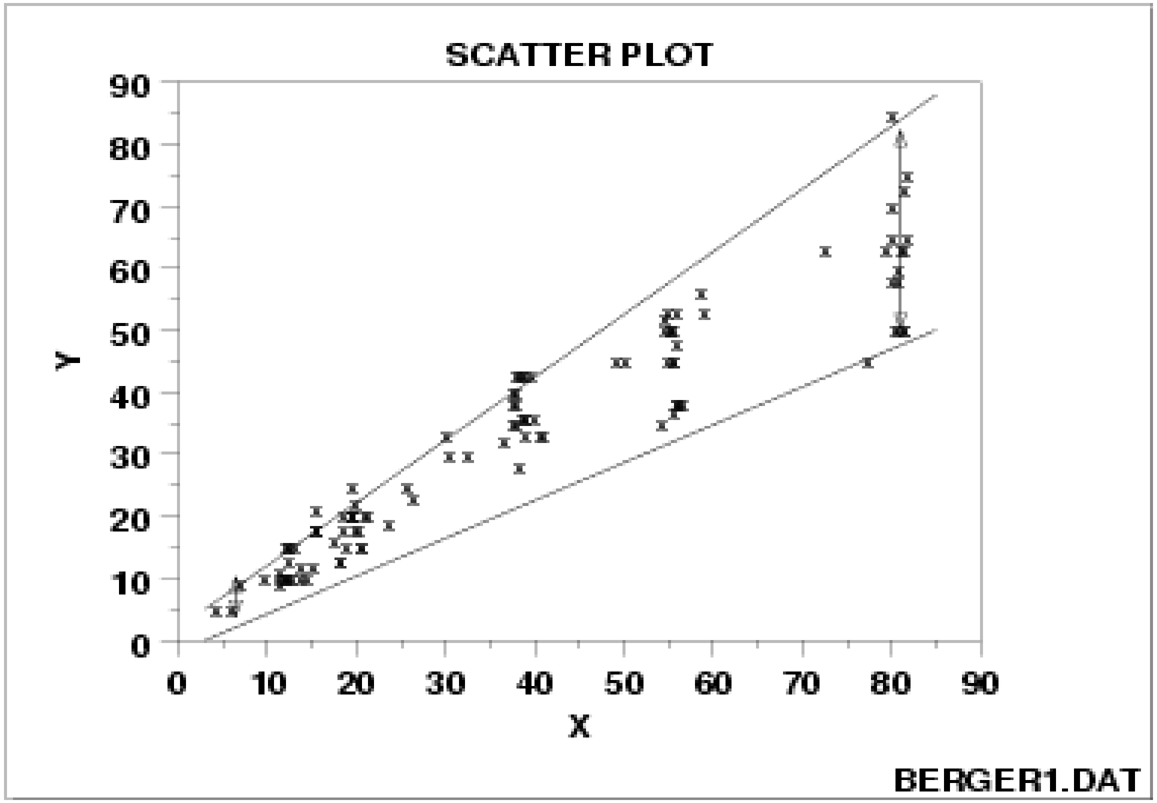


Scatter plot: Homoscedastic



Why is this important in classical statistical modelling?

Scatter plot: Heteroscedastic



variation in *Y* differs depending on the value of *X*

e.g., Y = annual tax paid, X = income

Two variables - continuous

- Scatterplots
  - But can be bad with lots of data

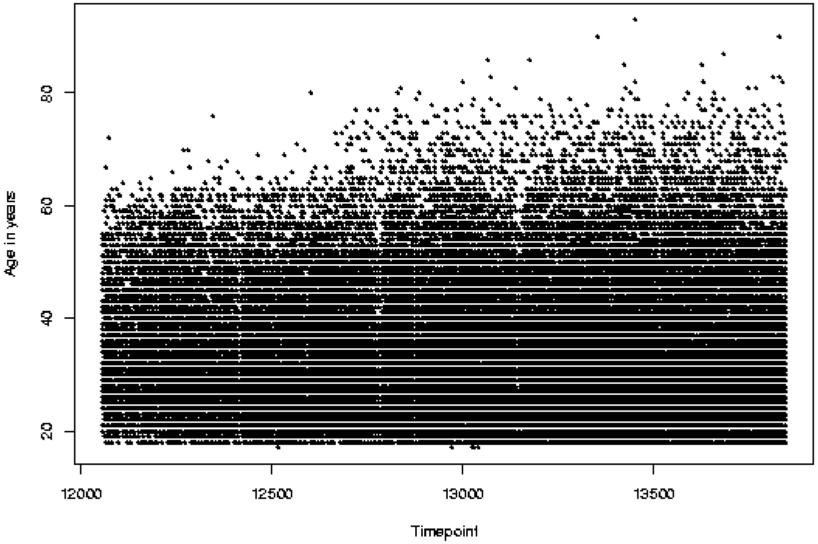
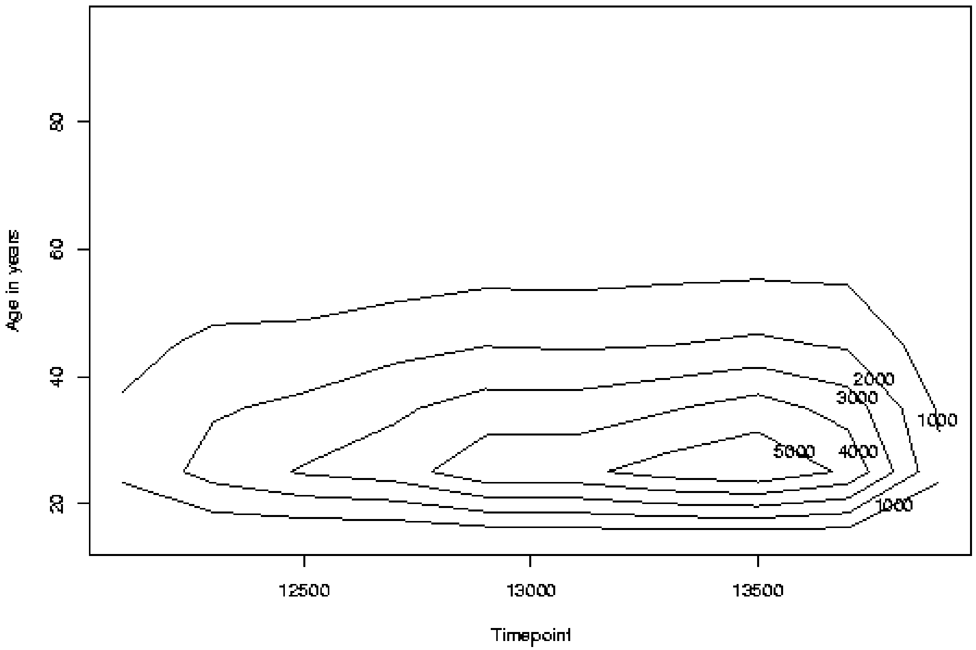


Figure 3.7: A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.

Two variables - continuous

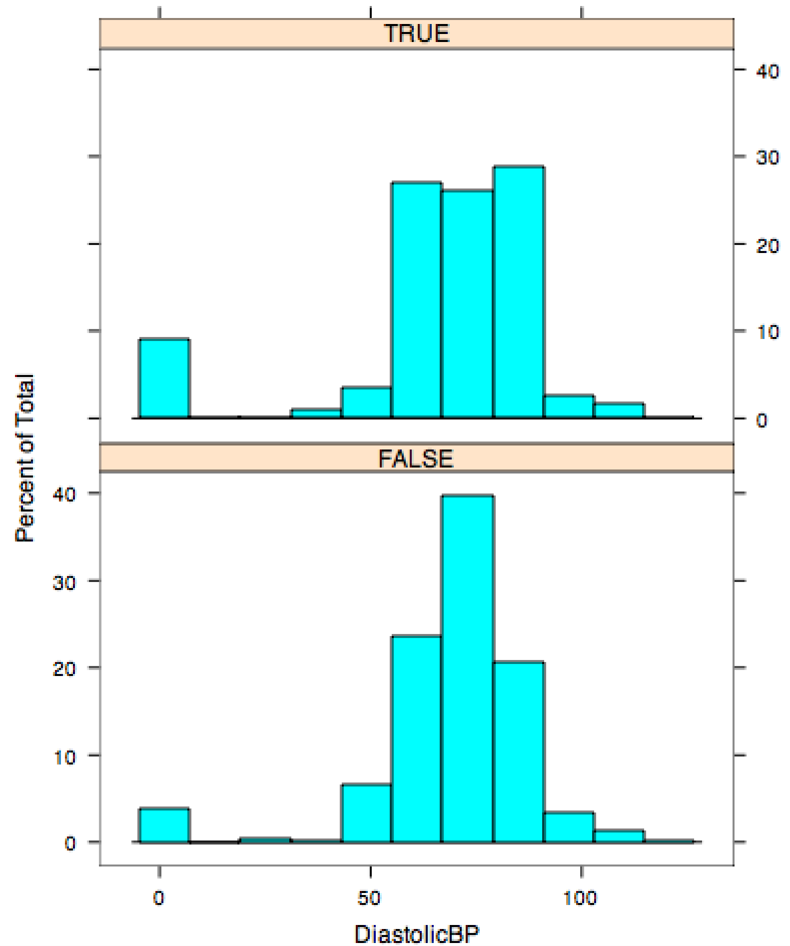
- What to do for large data sets
  - Contour plots



Displaying Two Variables

If one variable is categorical, use small multiples

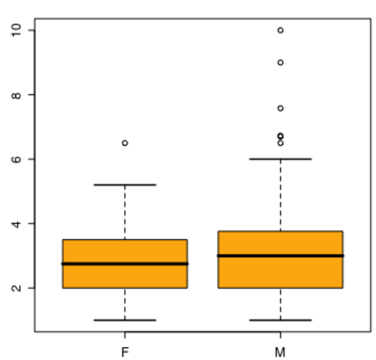
Many software packages have this implemented as ‘lattice’ or ‘trellis’ packages

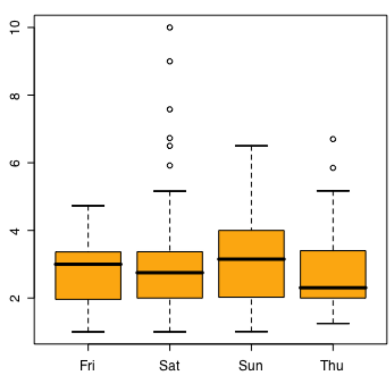


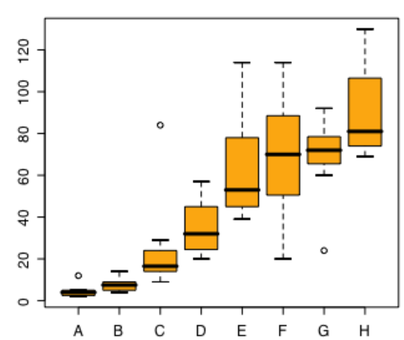
library(‘lattice’)
histogram(~DiastolicBP | TimesPregnant==0)

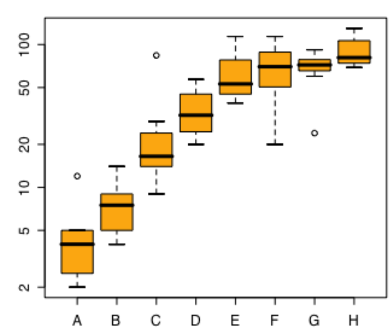
Two Variables - one categorical

- 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



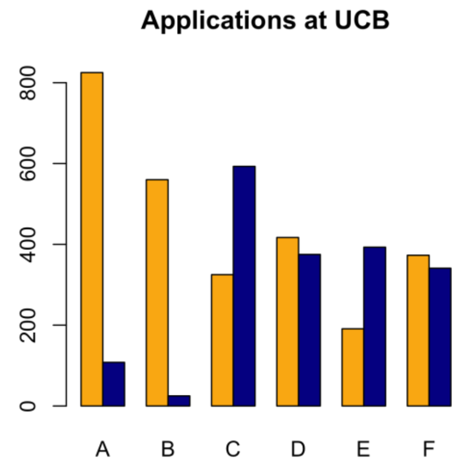


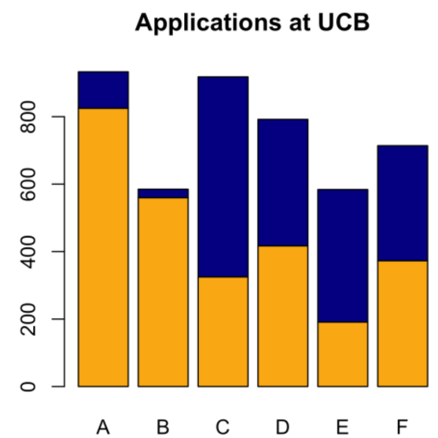




Barcharts and Spineplots

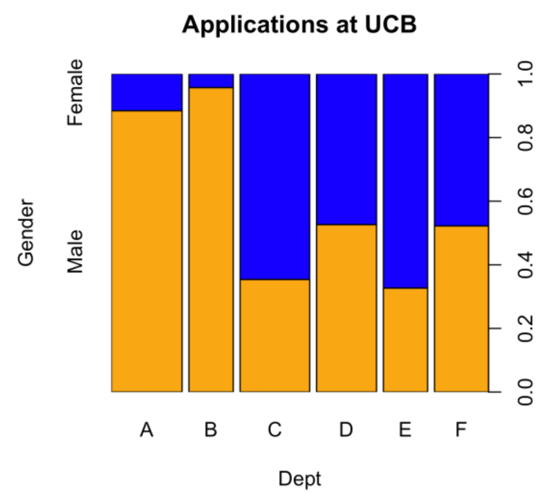
stacked barcharts can be used to compare continuous values across two or more categorical ones.





orange=M
blue=F

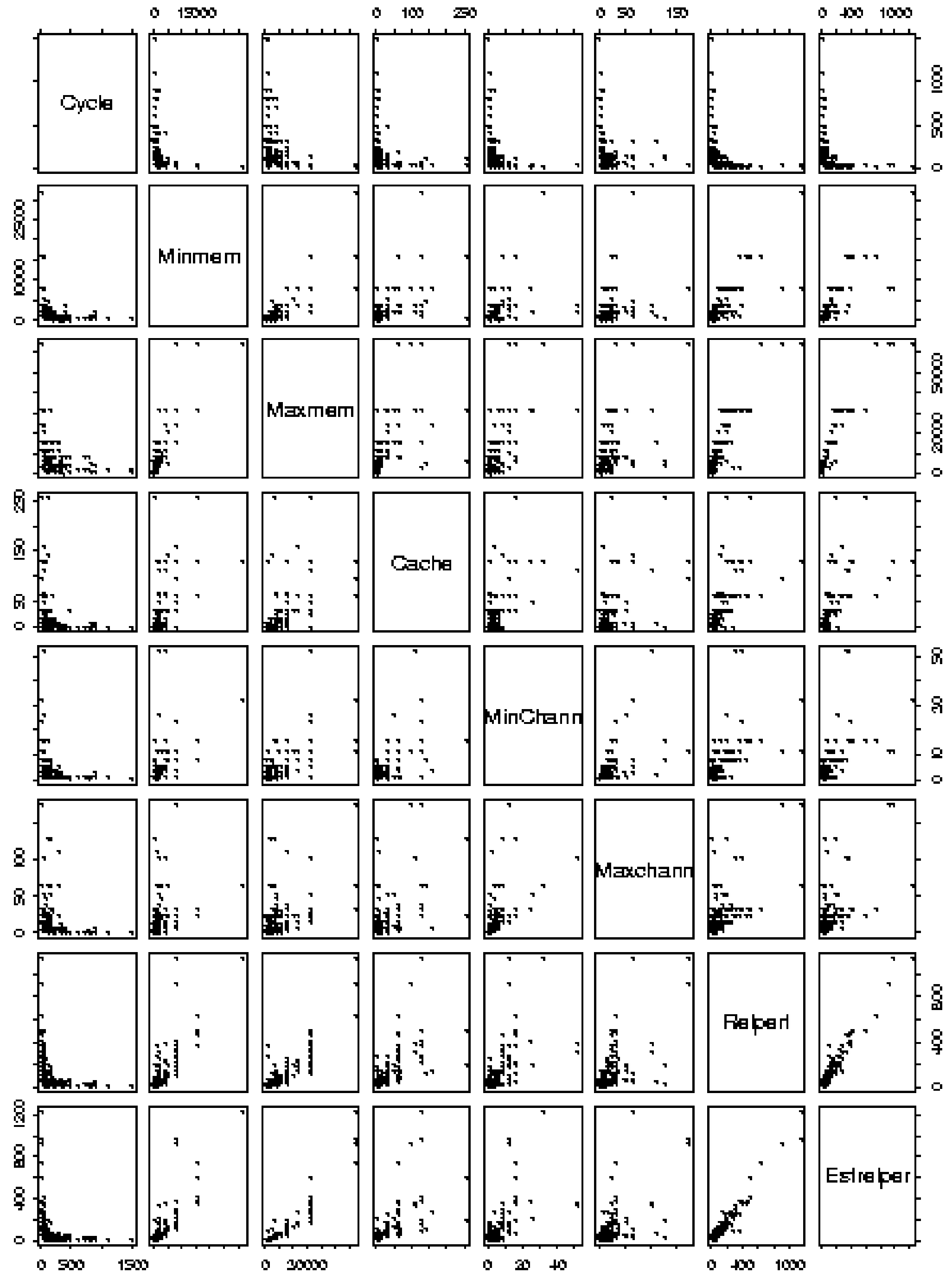
spineplots show proportions well, but can be hard to interpret

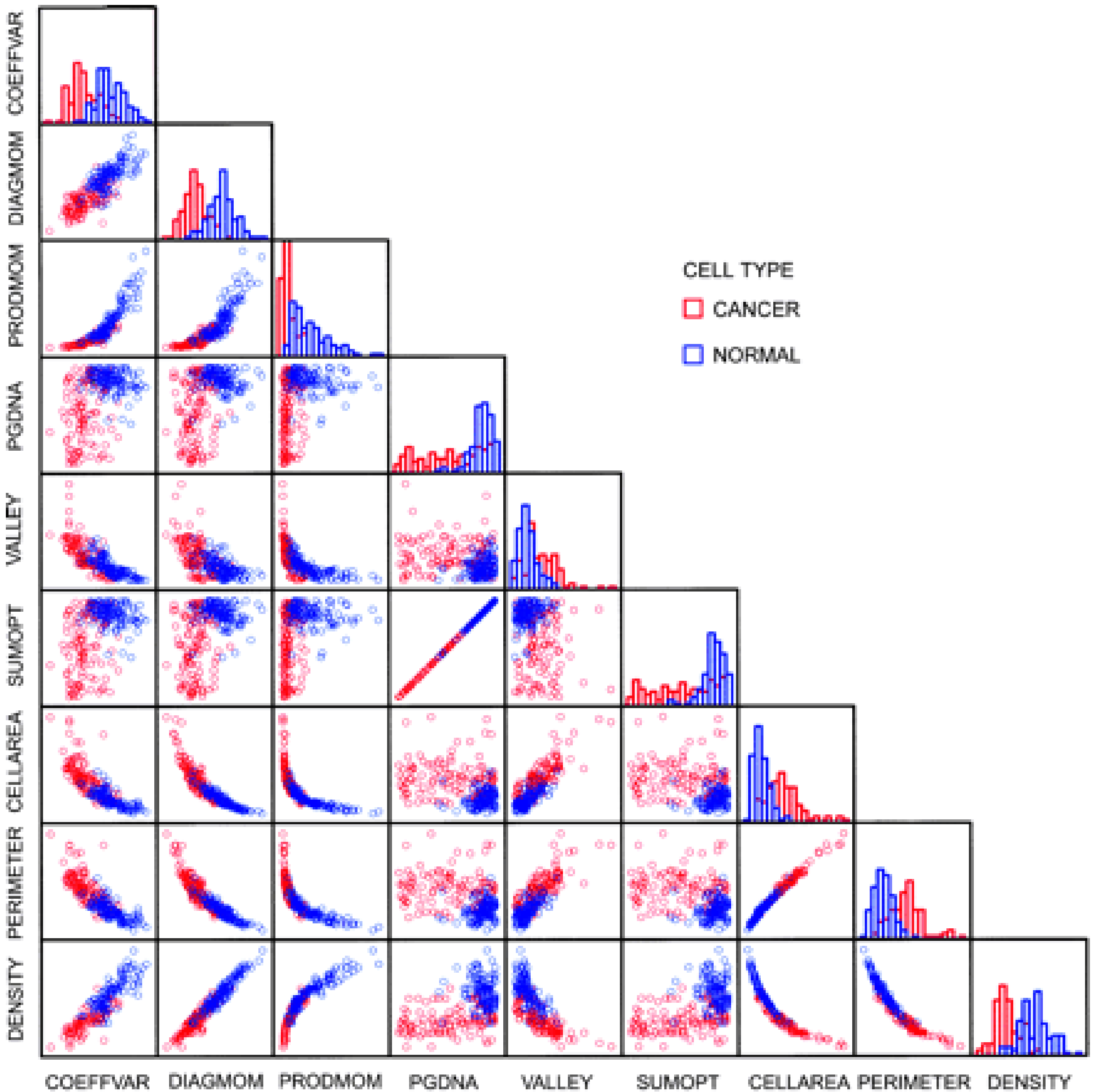


More than two variables

Pairwise scatterplots

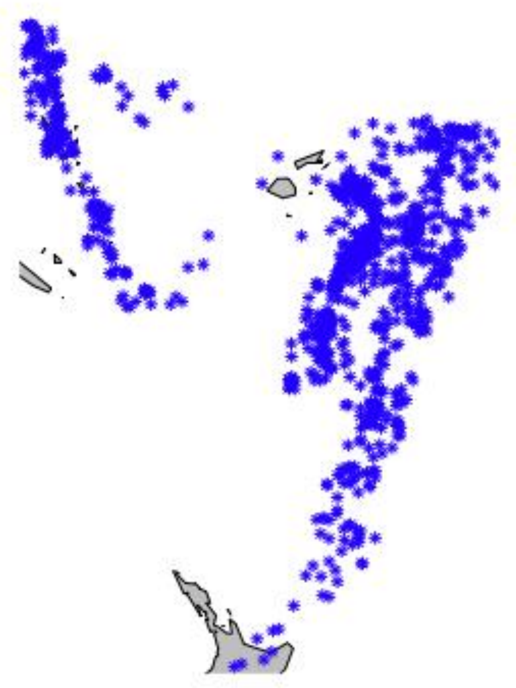
Can be somewhat ineffective for categorical data



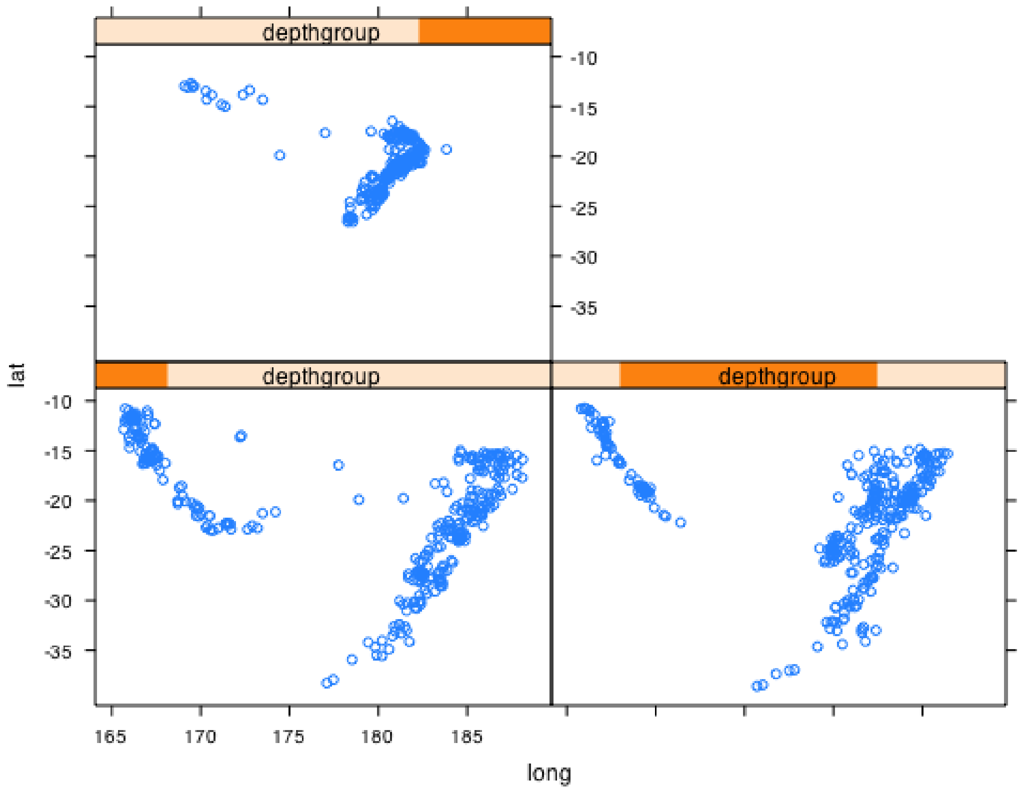

CELL TYPE
CANCER
NORMAL
COEFFVAR
DIAGMOM
PRODMOM
PGDNA
VALLEY
SUMOPT
CELLAREA
PERIMETER
DENSITY

Multivariate: More than two variables

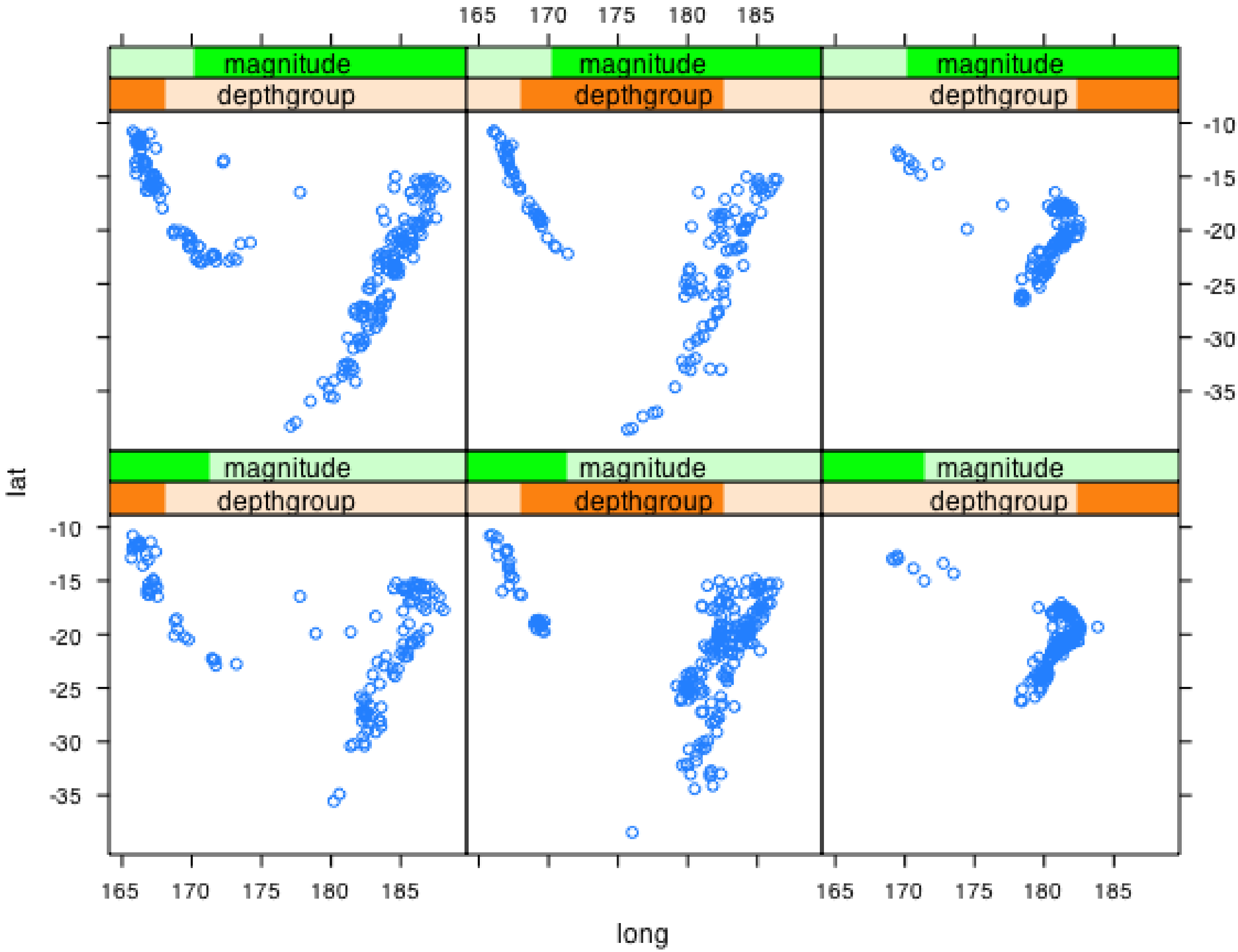
- Get creative!
- Conditioning on variables
  - trellis or lattice plots
  - Cleveland models on human perception, all based on conditioning
  - Infinite possibilities



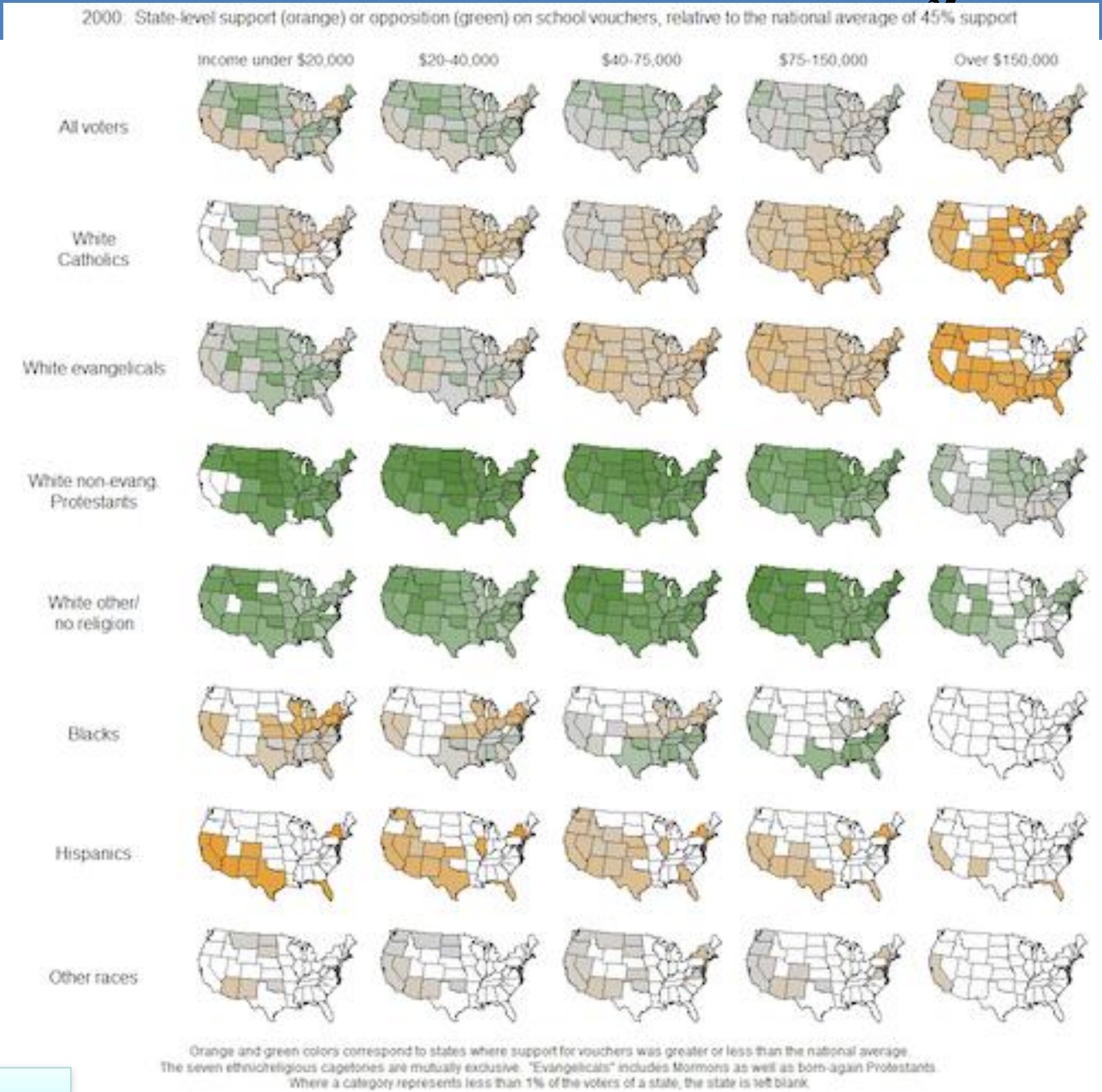
- Earthquake data:
  - locations of 1000 seismic events of MB > 4.0. The events occurred in a cube near Fiji since 1964
  - Data collected on the severity of the earthquake



depthgroup
depthgroup
depthgroup
-10
-15
-20
-25
-30
-35
lat
165
170
175
180
185
long

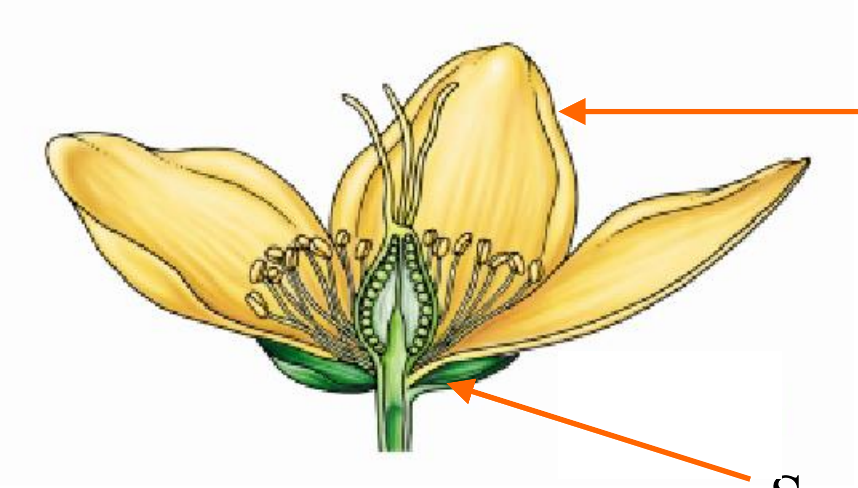

magnitude
depthgroup
lat
long
165
170
175
180
185
-10
-15
-20
-25
-30
-35

How many dimensions are represented here?



Andrew Gelman blog 7/15/2009

Multivariate Vis: Parallel Coordinates

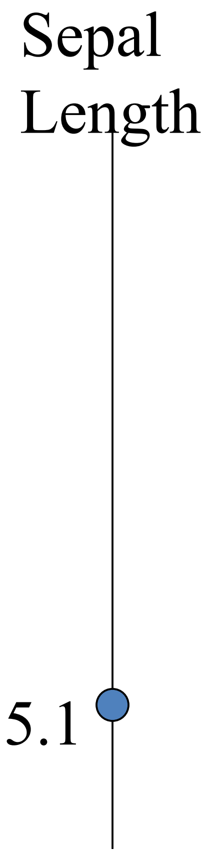


Petal, a non-reproductive part of the flower

Sepal, a non-reproductive part of the flower

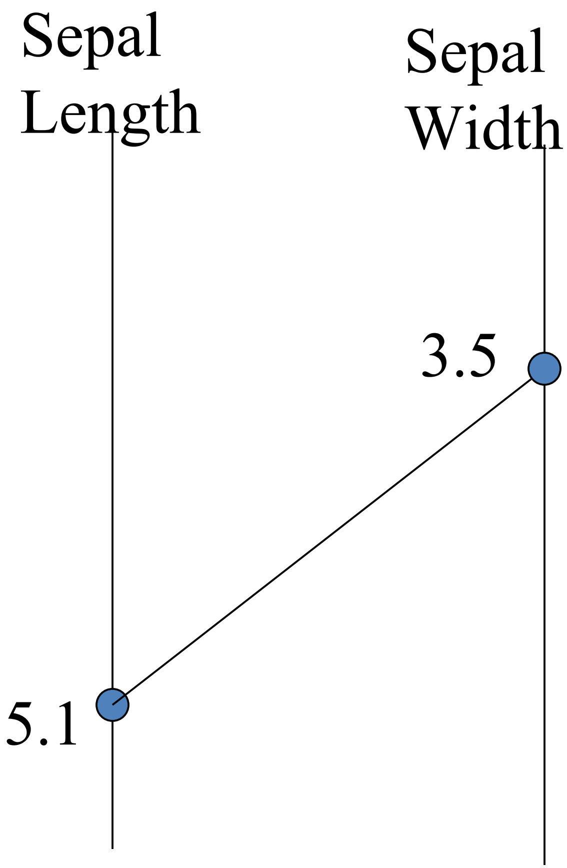
The famous iris data!

Parallel Coordinates



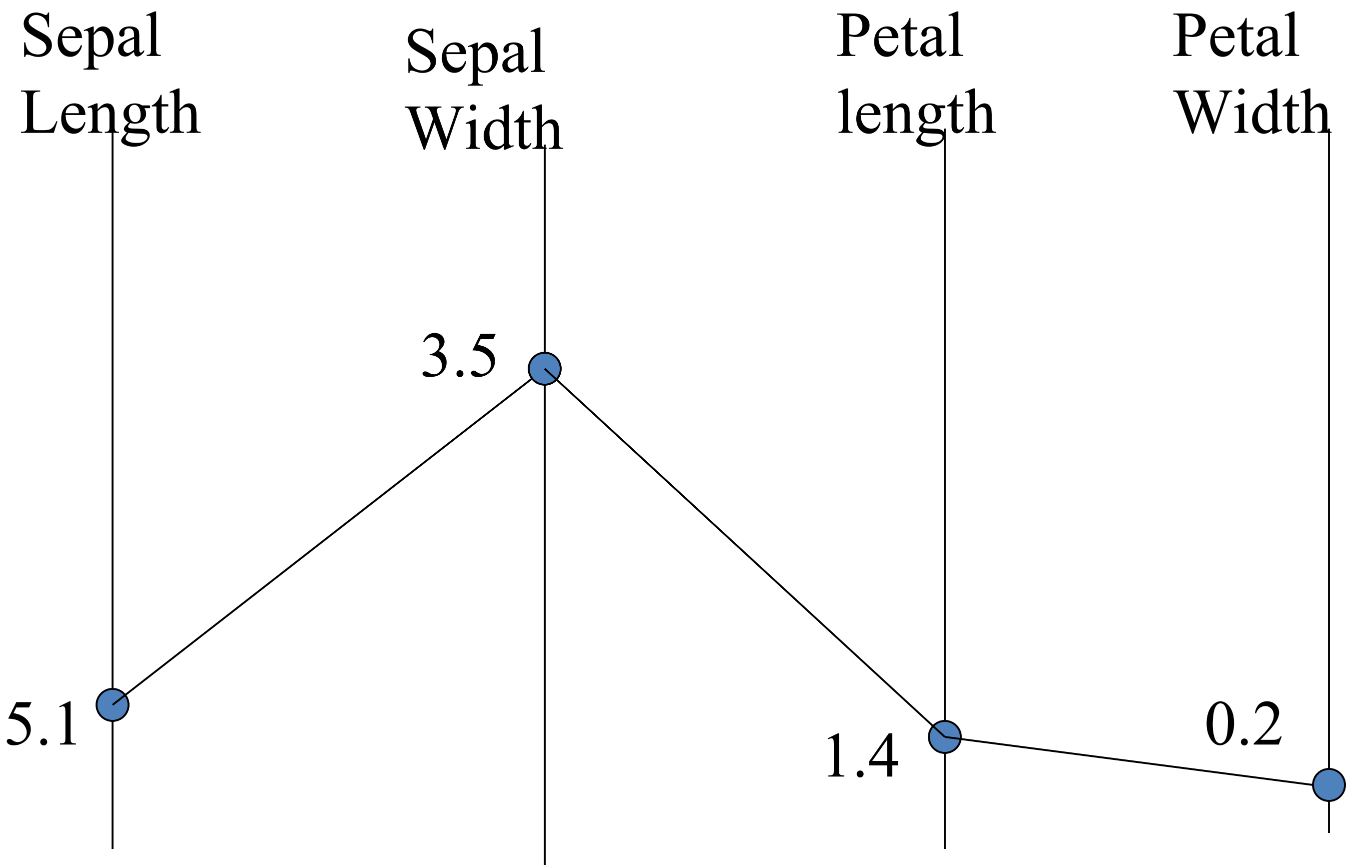
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 2 D



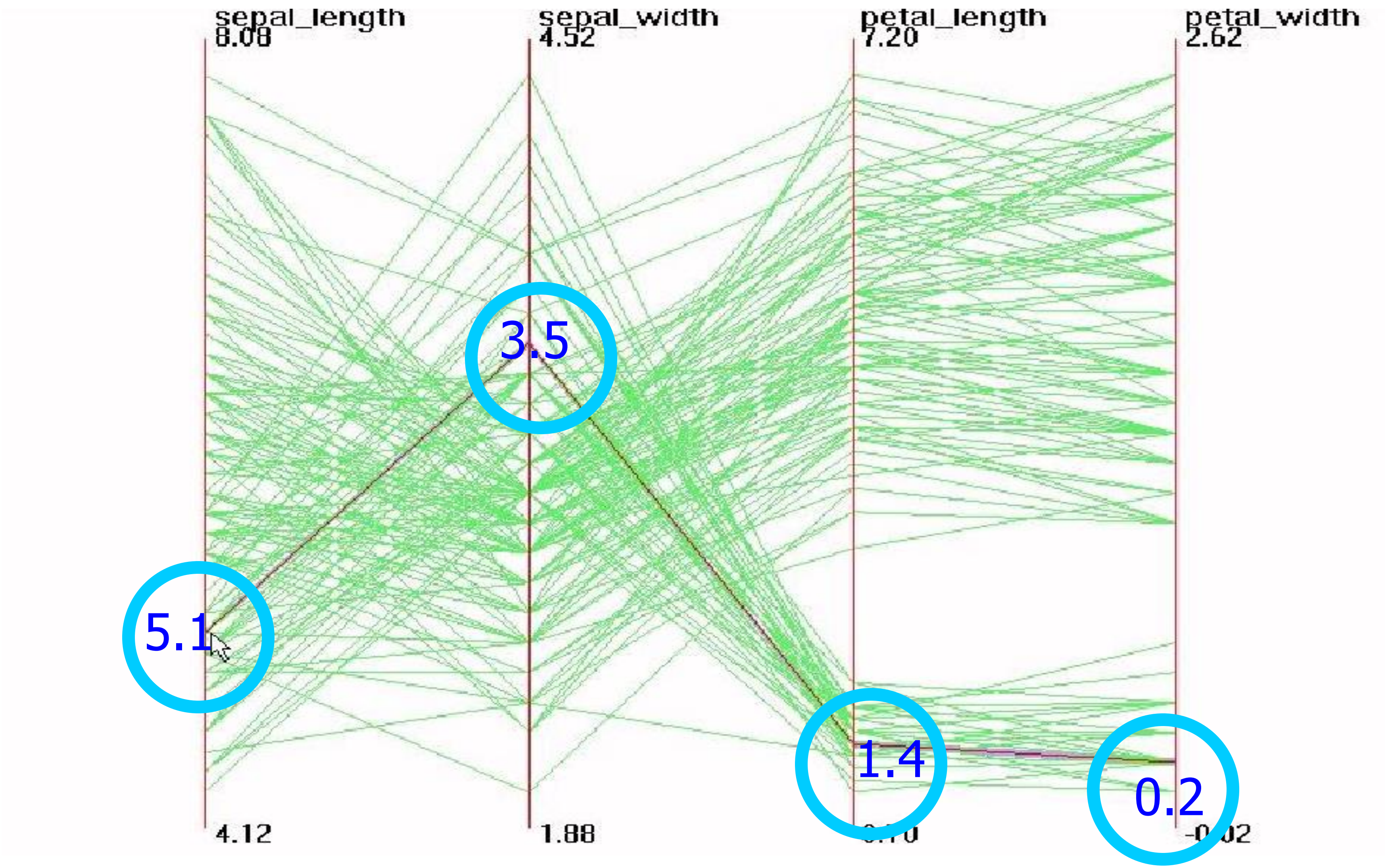
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 4 D

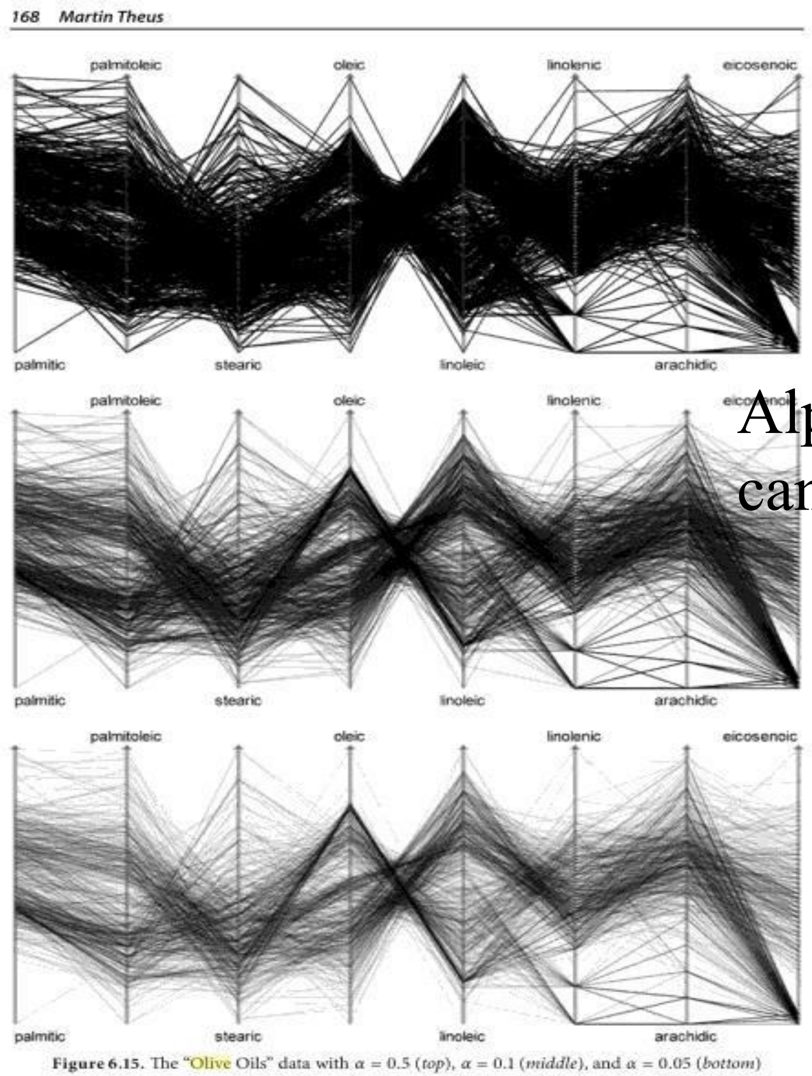


sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Visualization of Iris data



Multivariate:

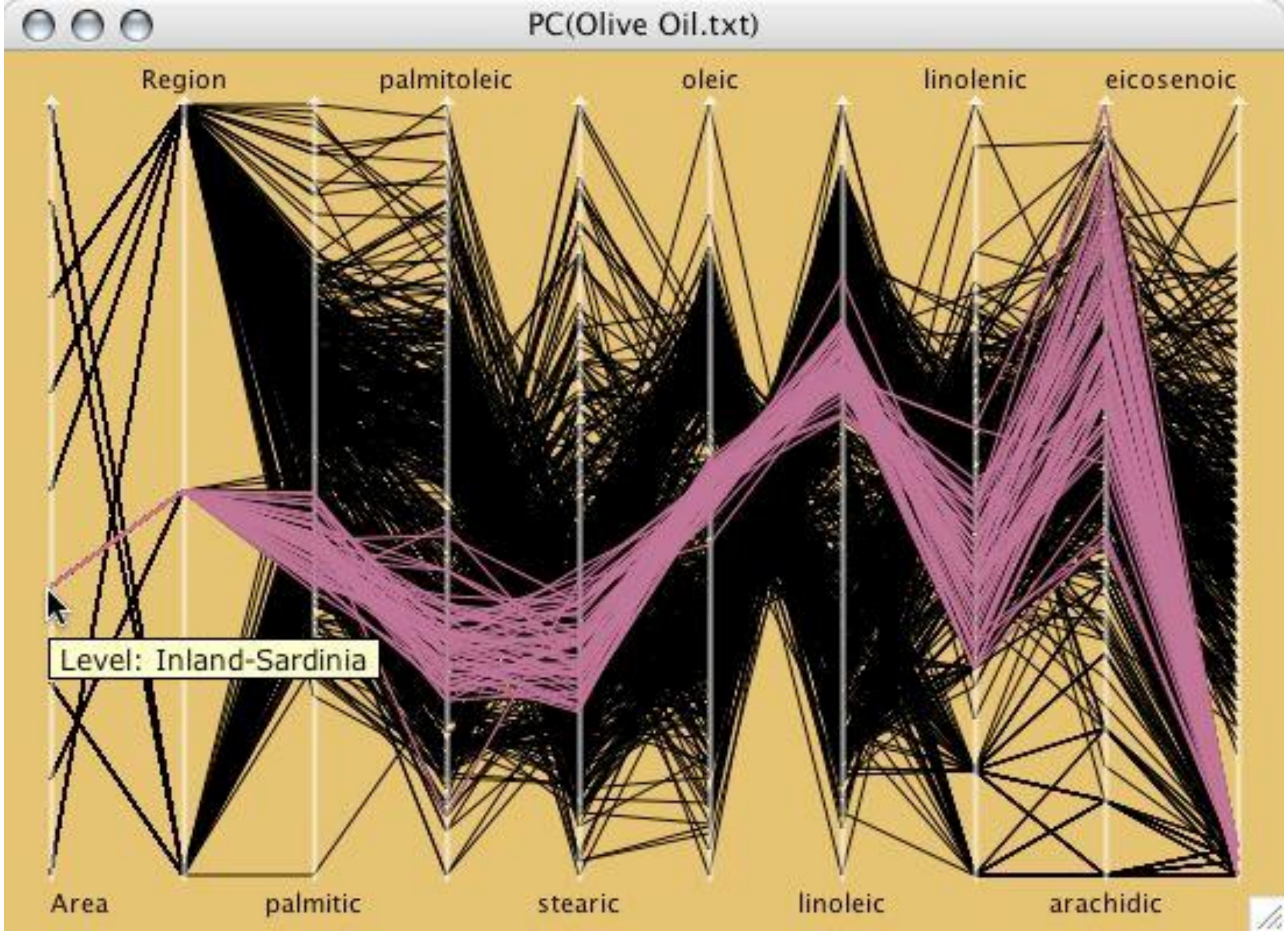


Alpha blending can be effective

Courtesy Unwin, Theus, Hofmann

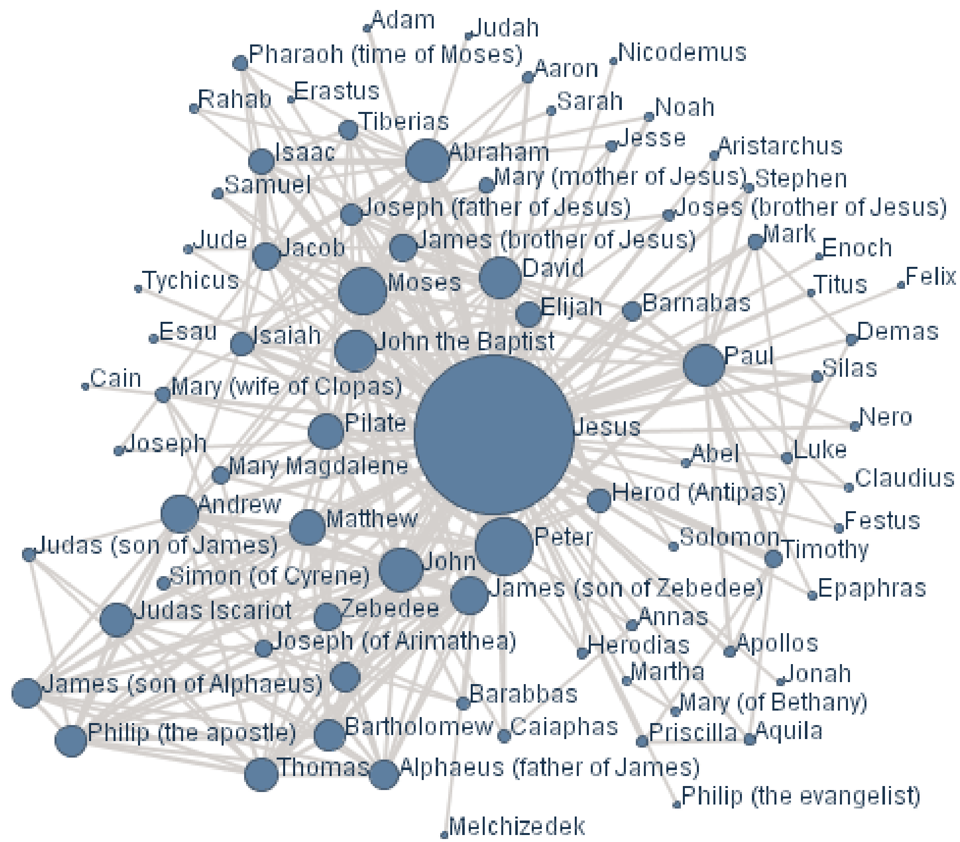
Parallel coordinates

Useful in an interactive setting



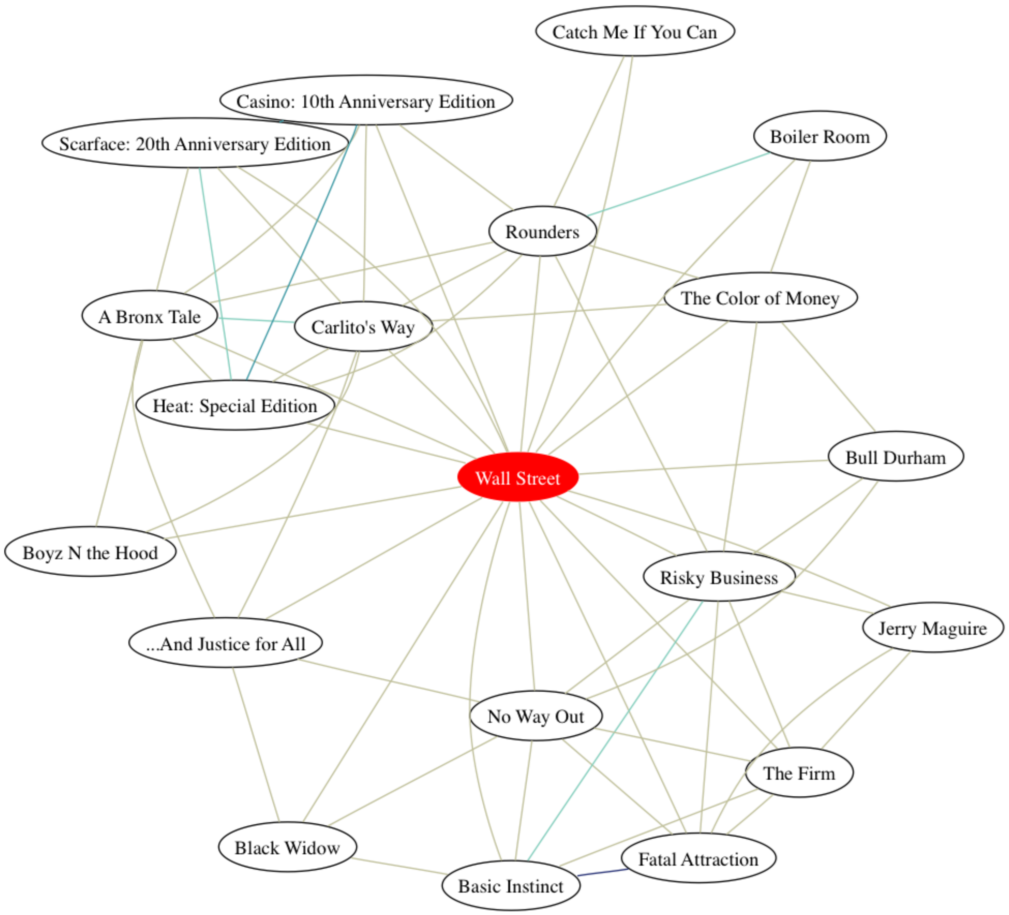
Networks and Graphs

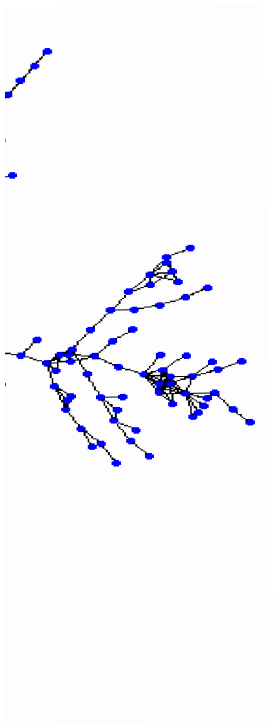
Visualizing networks is helpful, even if is not obvious that a network exists



Network Visualization

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

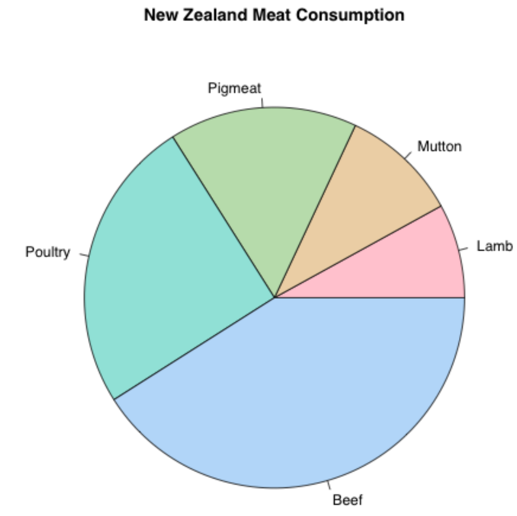




What’s missing?

pie charts

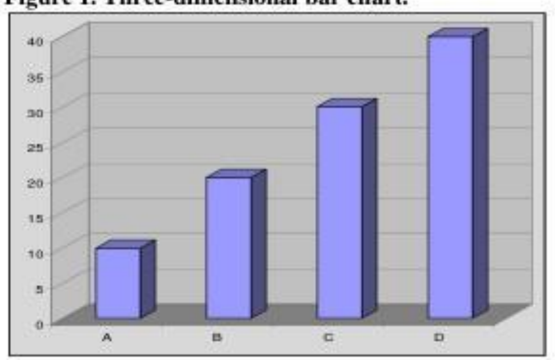
- very popular
- good for showing simple relations of proportions
- Human perception not good at comparing arcs
- barplots, histograms usually better (but less pretty

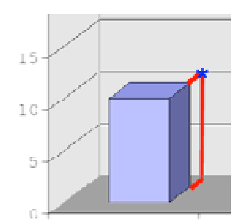


3D

- 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/

Figure 1. Three-dimensional bar chart.





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

More on this when we talk about Data Visualization

Visualization done right

- Hans Rosling @ TED
- http://www.youtube.com/watch?v=jbkSRLYSojo