CMSC828G

Principles of Data Mining

Lecture #3

- Today's Reading:
  - HMS, chapter 3
- Today's Lecture:
  - Exploratory data analysis
  - Statistical Graphics
  - Data reduction
    - PCA
    - multidimensional scaling
- Upcoming Due Dates:
  - P0 due 2/7

Exploratory Data Analysis (EDA)

- An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to
  1. maximize insight into a data set;
  2. uncover underlying structure;
  3. extract important variables;
  4. detect outliers and anomalies;
  5. test underlying assumptions;
  6. develop parsimonious models; and
  7. determine optimal factor settings

NIST/Semantech Engineering Statistics Handbook

EDA cont.

- data driven hypothesis generation
- interactive, visual
- Makes use of human capacity for pattern recognition
  - limited to three to five dimensions – natural senses
  - ten if we count derivatives
- Techniques primarily graphical, statistical graphics
  - plotting raw data
  - plotting simple statistics
  - data reduction for high dimensional data

vs. statistical data analysis (SDA), i.e. hypothesis testing

Plotting Raw Data (1D)

- Data traces
- Histograms

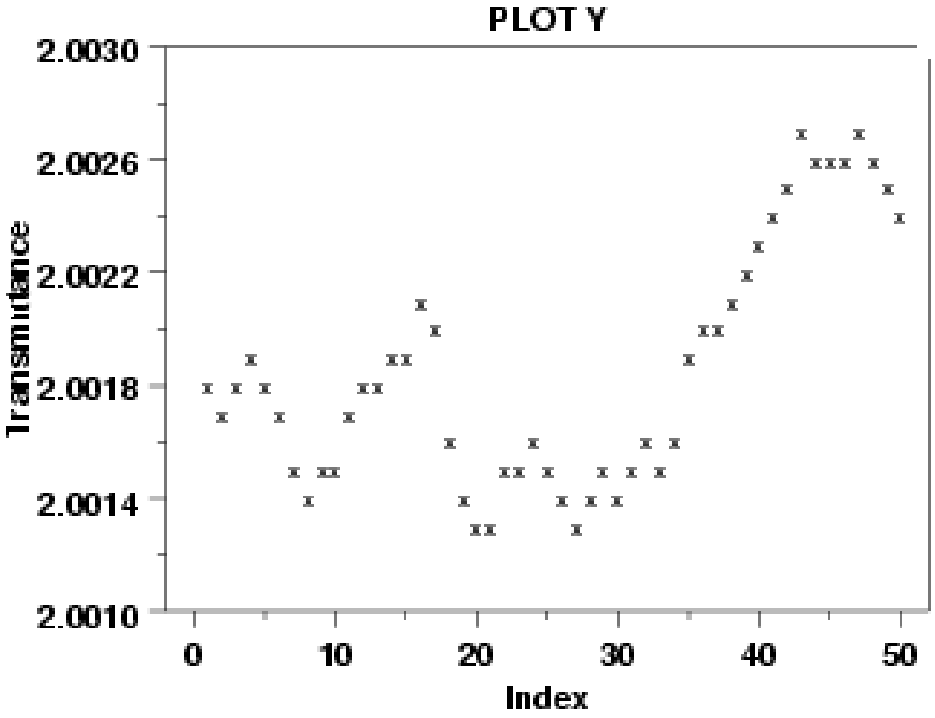
Run Sequence Plot

- Purpose:
  check for shifts in location and scale and outliers

- Run sequence plots formed by:
  Vertical axis:
    Response variable Y(i)
  Horizontal axis:
    Index i (i = 1, 2, 3, ...)

- The run sequence plot can be used to answer the following questions

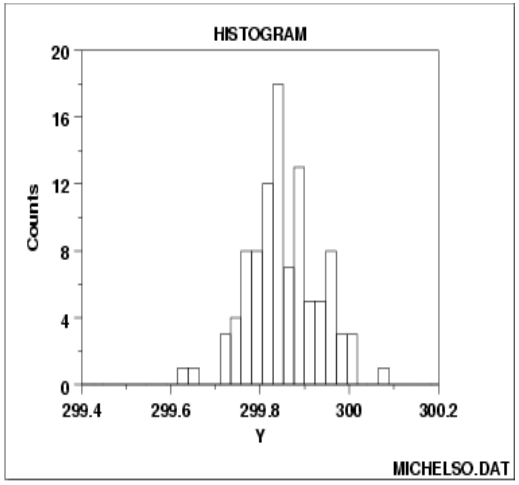
  1. Are there any shifts in location?
  2. Are there any shifts in variation?
  3. Are there any outliers?



Histogram

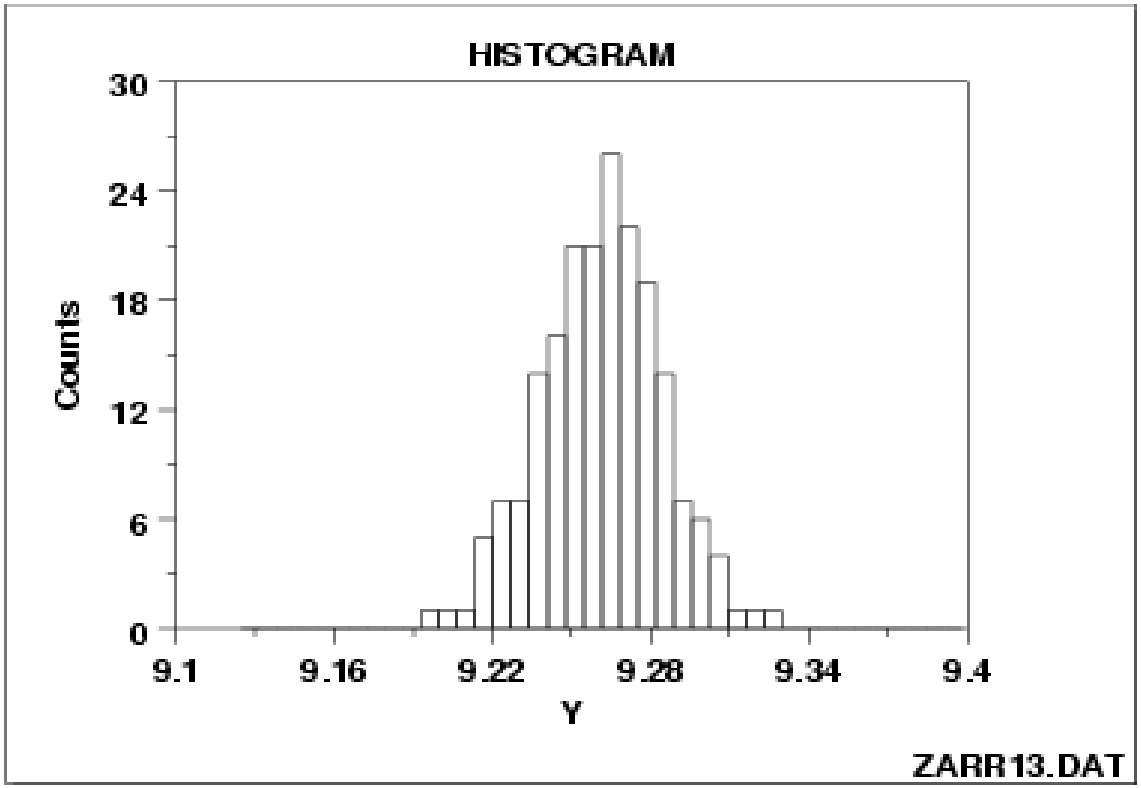
- Most common form: split data range into equal-sized bins Then for each bin, count the number of points from the data set that fall into the bin.
  - Vertical axis: Frequency (i.e., counts for each bin)
  - Horizontal axis: Response variable

- The histogram graphically shows the following:
  1. center (i.e., the location) of the data;
  2. spread (i.e., the scale) of the data;
  3. skewness of the data;
  4. presence of outliers; and
  5. presence of multiple modes in the data.



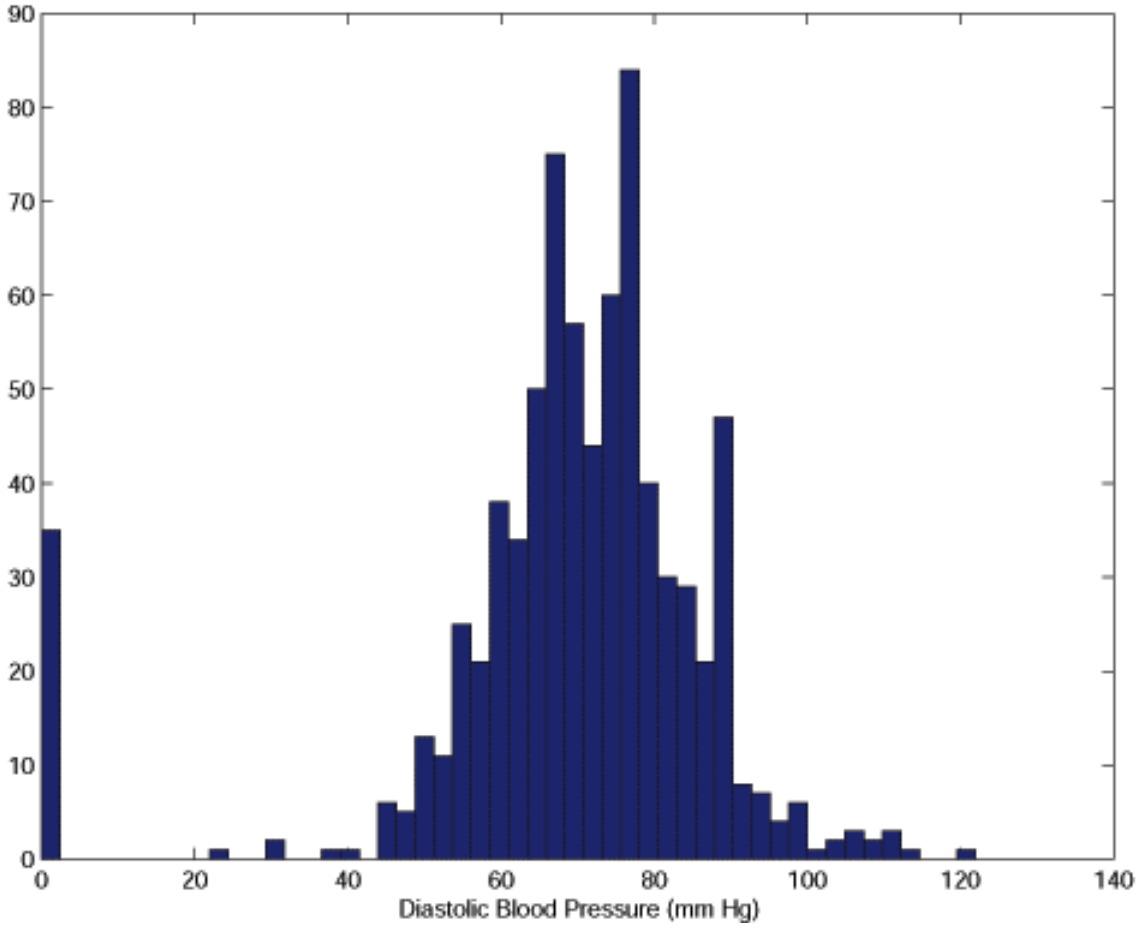
These features provide strong indications of the proper distributional model for the data

Histogram Example



classical bell-shaped, symmetric histogram with most of the frequency counts bunched in the middle and with the counts dying off out in the tails. From a physical science/engineering point of view, the normal distribution is that distribution which occurs most often in nature (due in part to the central limit theorem).

Another Histogram Example



Taxonomy of Histograms

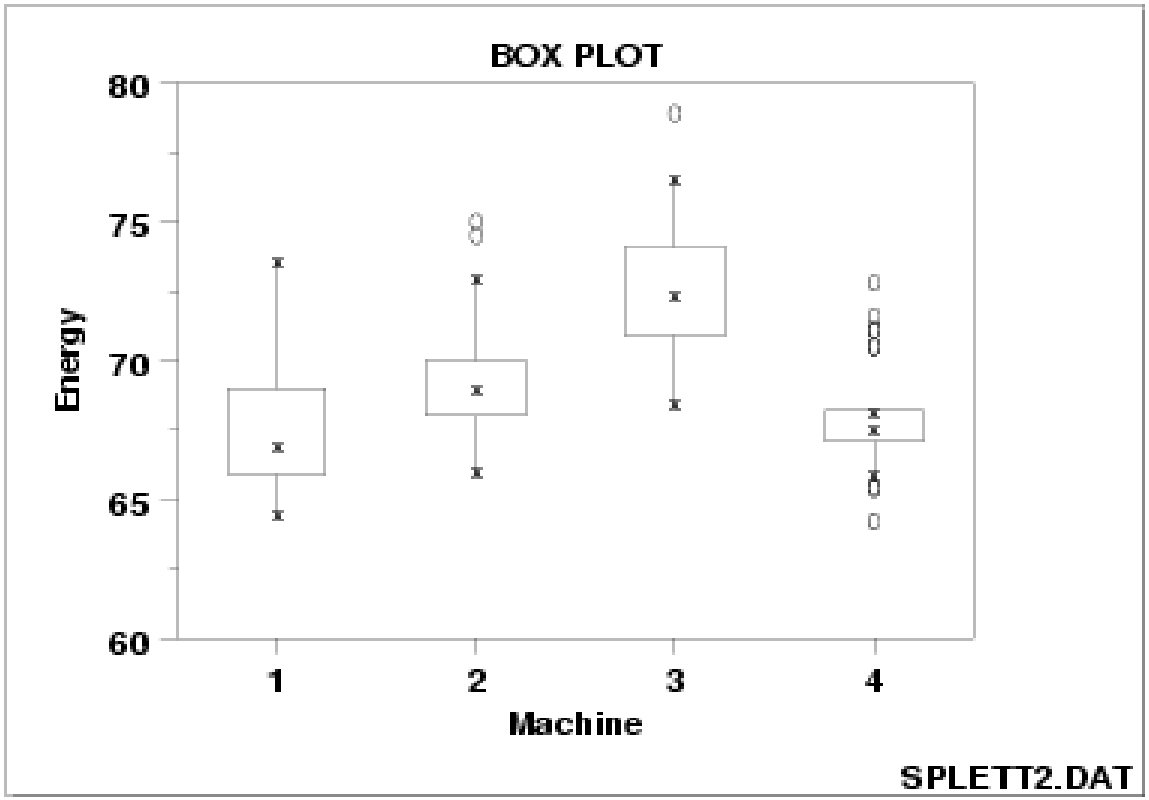
- equiwidth – each bucket sub-range is the same size
- equidepth – the number of data points in each bucket is the same
- others?

Histograms cont.

- For small data sets, histograms can be misleading. Small changes in the data or to the bucket boundaries can result in very different histograms.
- For large data sets, histograms can be quite effective at illustrating properties of the distribution.
- example

- Can smooth histogram using a variety of techniques
  - kernel estimates (we will discuss this in more detail when we discuss Kernel methods and SVMs)

Box Plot



1. Calculate the median and the quartiles
2. Plot the median and draw a box between lower and upper quartiles; this box represents the middle 50% of the data--the "body" of the data.
3. Draw a line from the lower quartile to the minimum point and another line from the upper quartile to the maximum point.

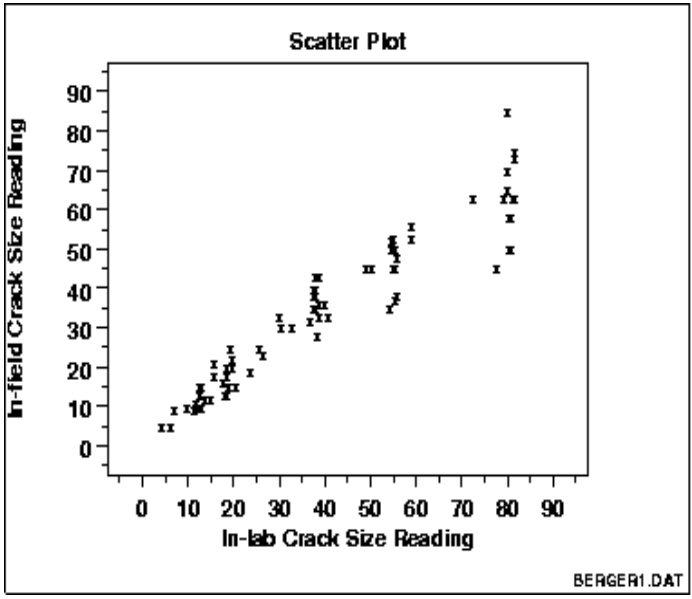
Summary Statistics

- measures of location
  - mean
  - median
  - mode – most common value
  - quartile – first quartile, value that is greater than ¼ of the data points; third quartile, value that is greater than ¾
- measures of dispersion or variability
  - standard deviation, variance
  - range – difference between the largest and smallest points
  - skewness – whether the distribution has a single long tail:

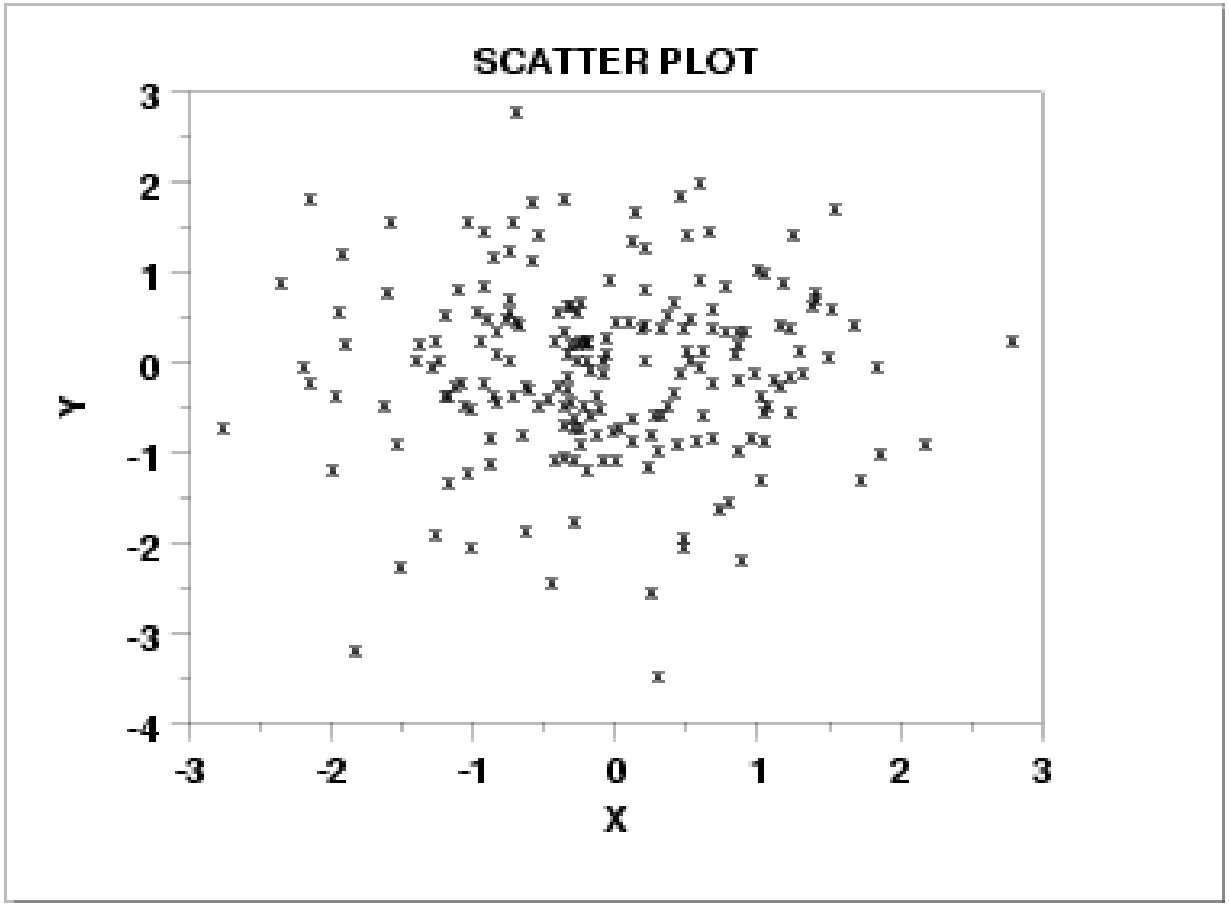
$$\frac{\sum \left(x(i) - \hat{\mu} \right)^3}{\left(\sum \left(x(i) - \hat{\mu} \right)^2 \right)^{\frac{3}{2}}}$$

2D: Scatter Plots

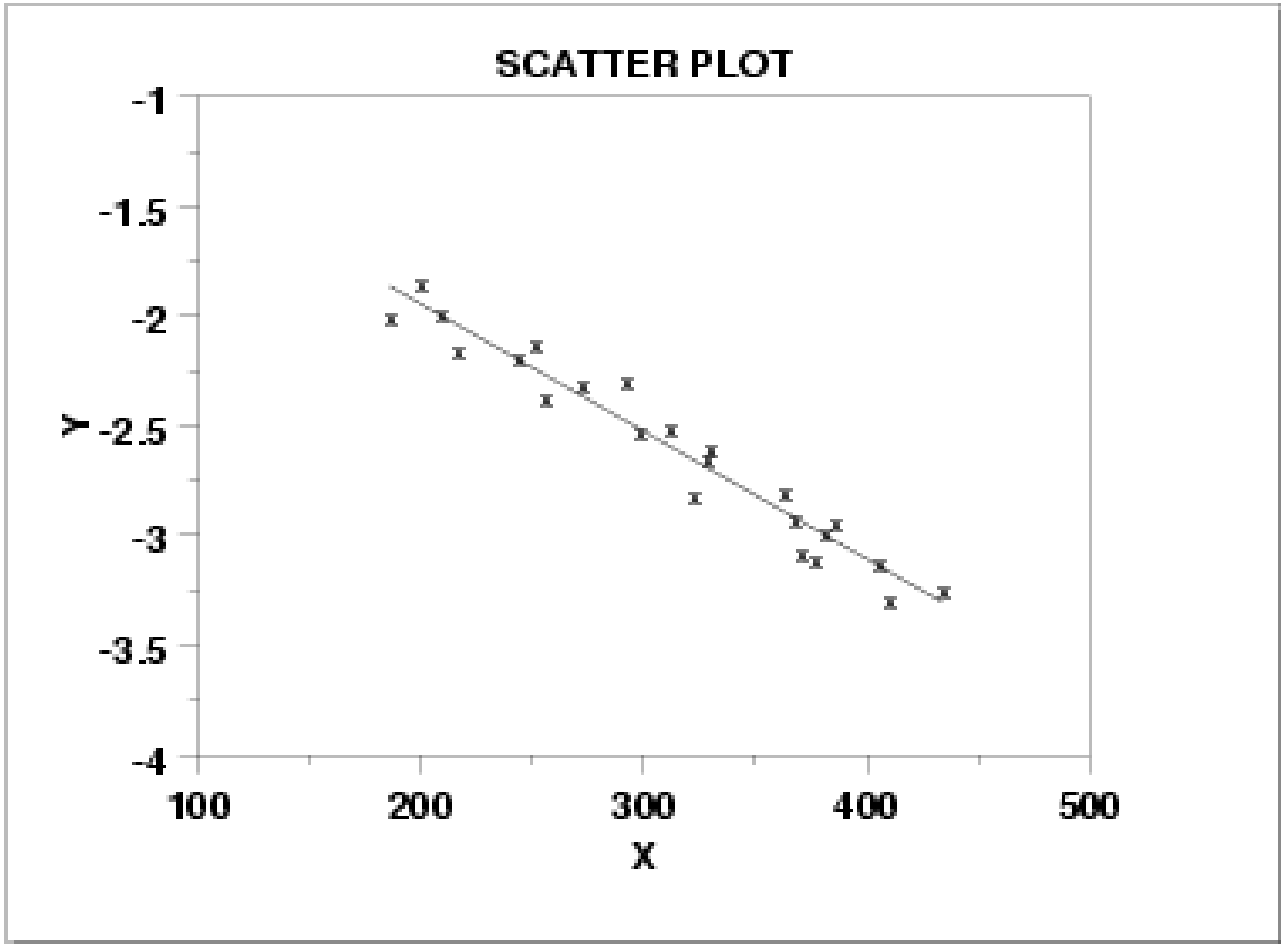
- standard tool for displaying relationship between two variables
- A scatter plot is a plot of the values of Y versus the corresponding values of X:
  - Vertical axis: variable Y--usually the response variable
  - Horizontal axis: variable X--variable we suspect may be related
- Scatter plots can provide answers to the following questions:
  1. Are variables X and Y related?
  2. Are variables X and Y linearly related?
  3. Are variables X and Y non-linearly related?
  4. Does the variation in Y change depending on X?
  5. Are there outliers?



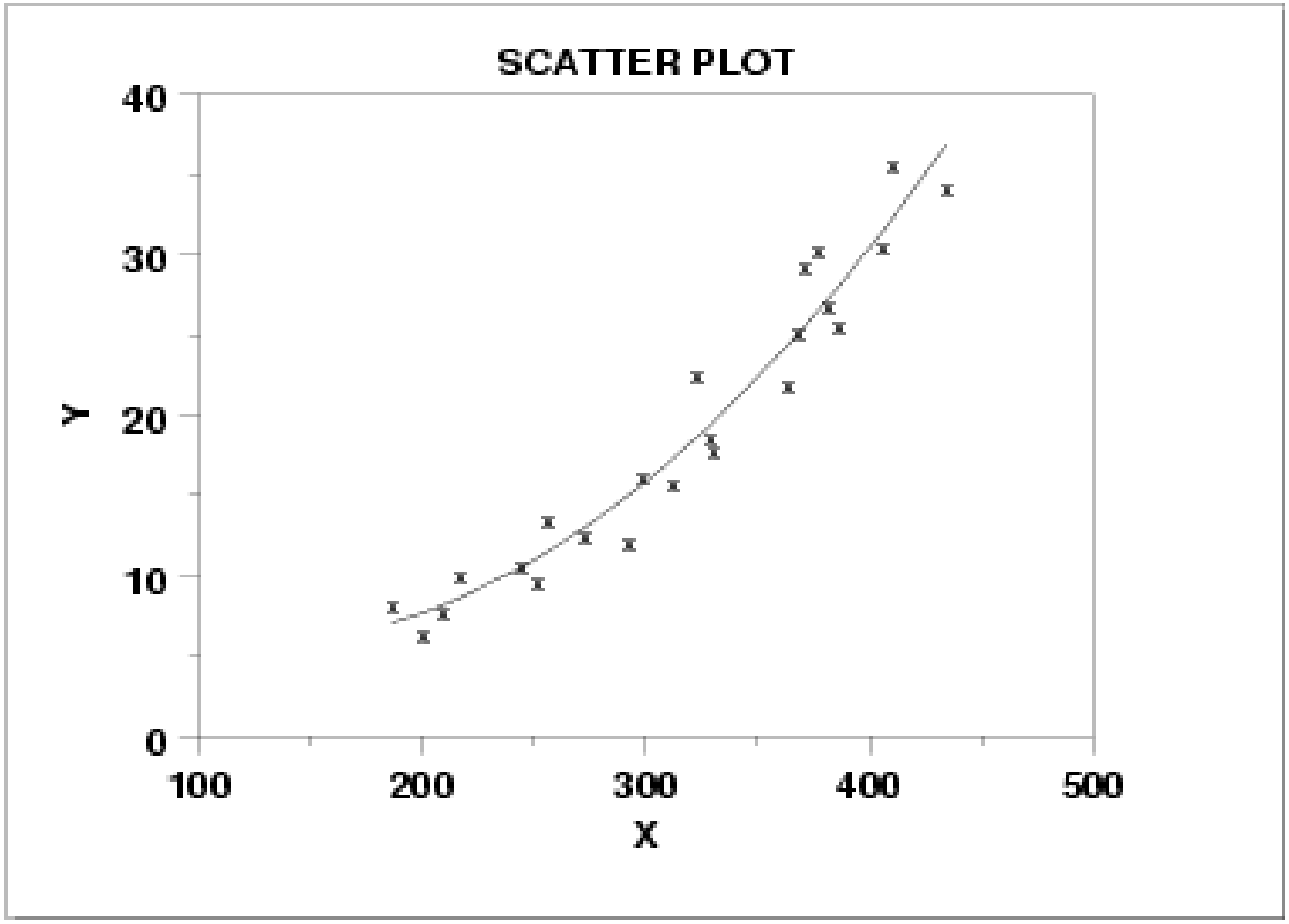
Scatter Plot: No relationship



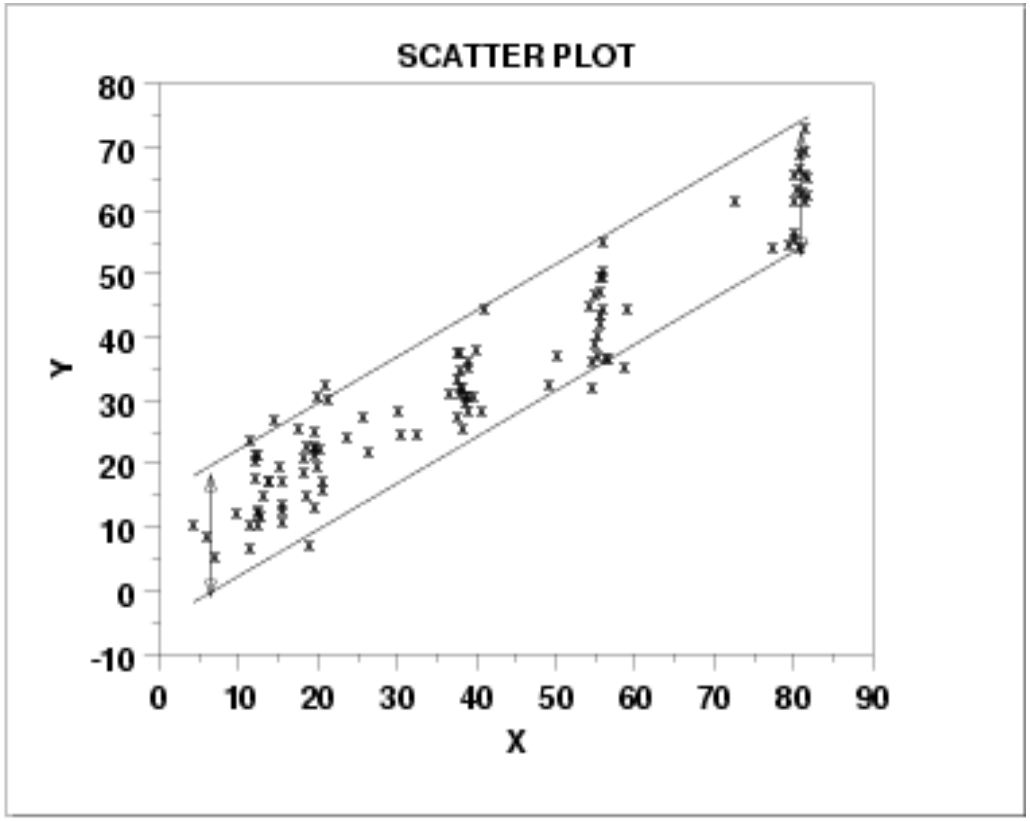
Scatter Plot: Linear relationship



Scatter Plot: Quadratic relationship

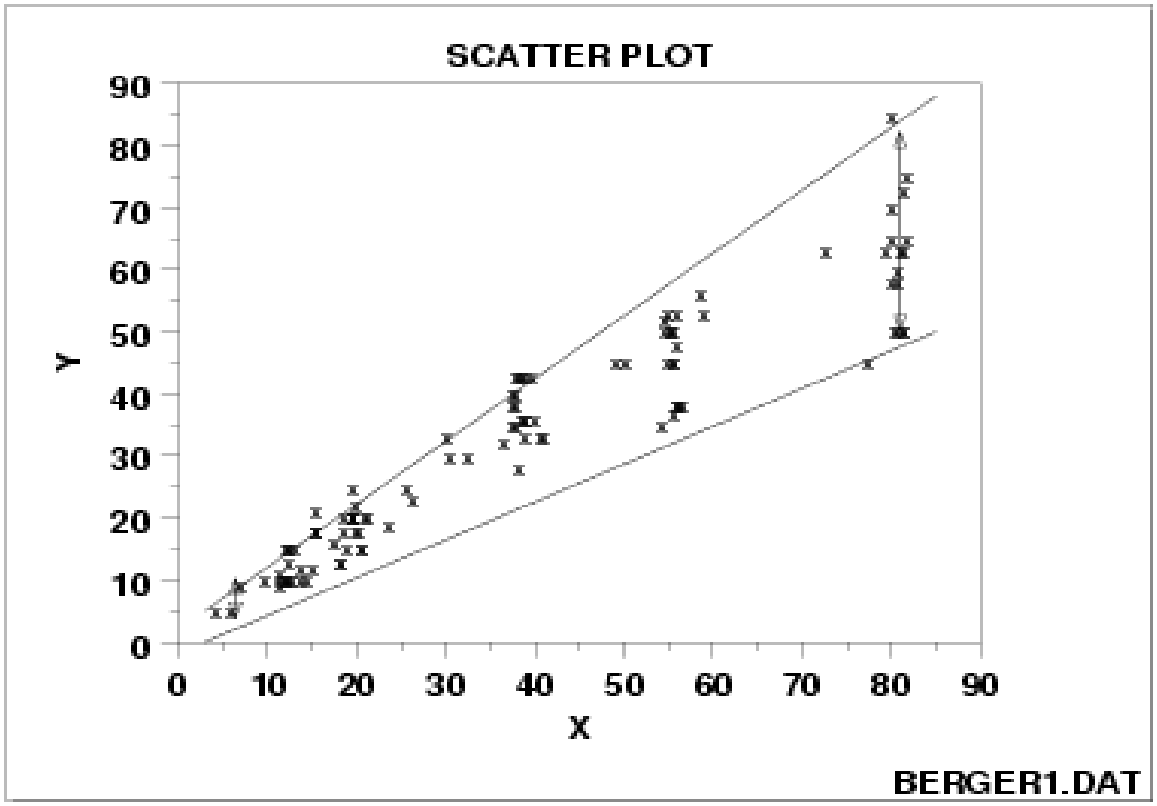


Scatter plot: Homoscedastic



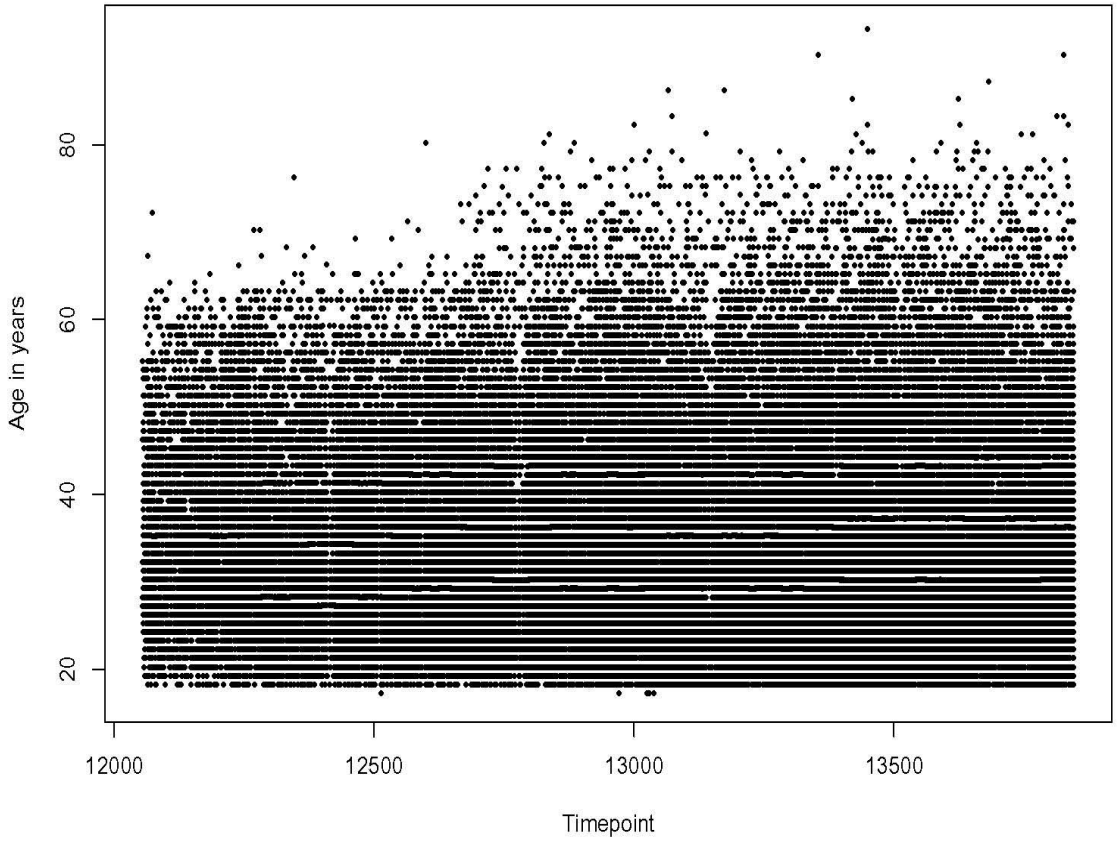
Variation of Y Does Not Depend on X

Scatter plot: Heteroscedastic



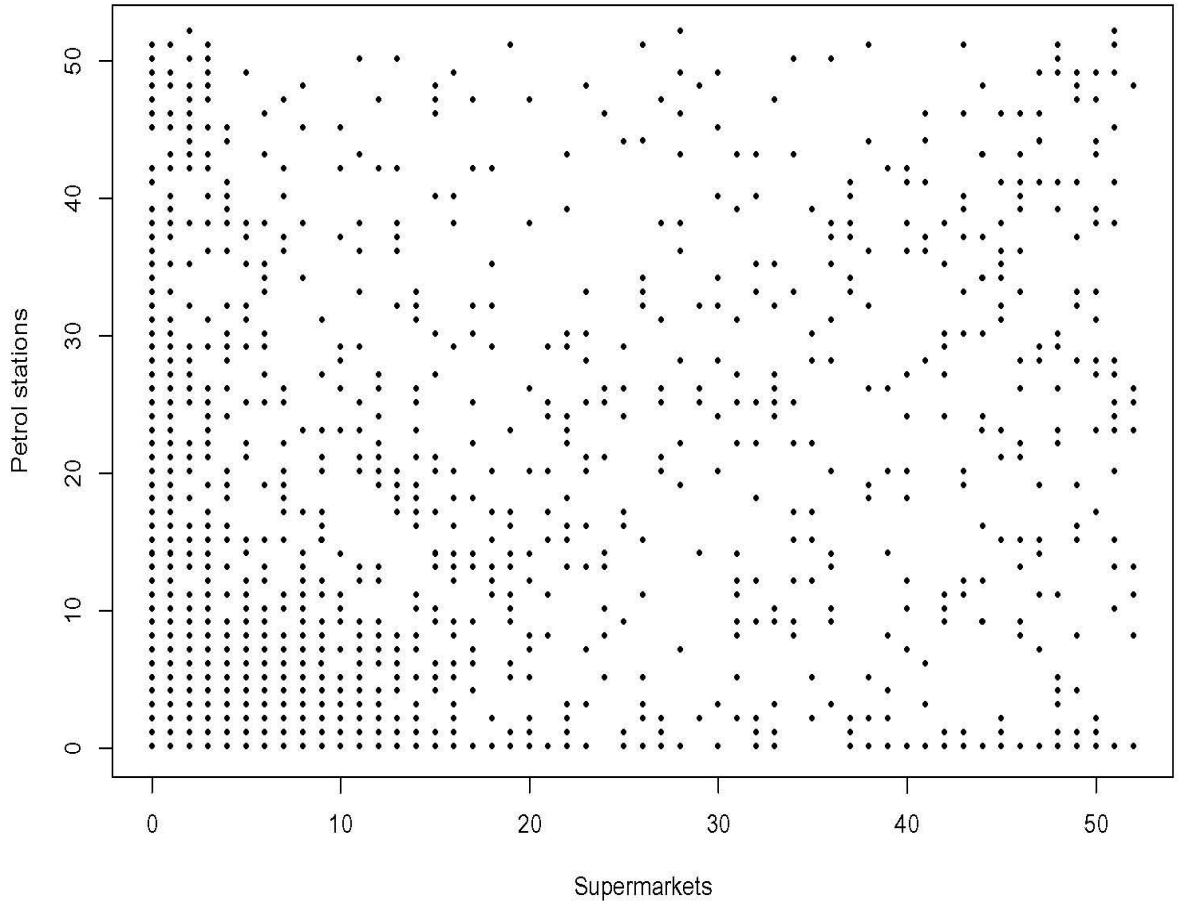
variation in *Y* differs depending on the value of *X*

Problems with scatter plots



too much data $\Rightarrow$ black rectangle

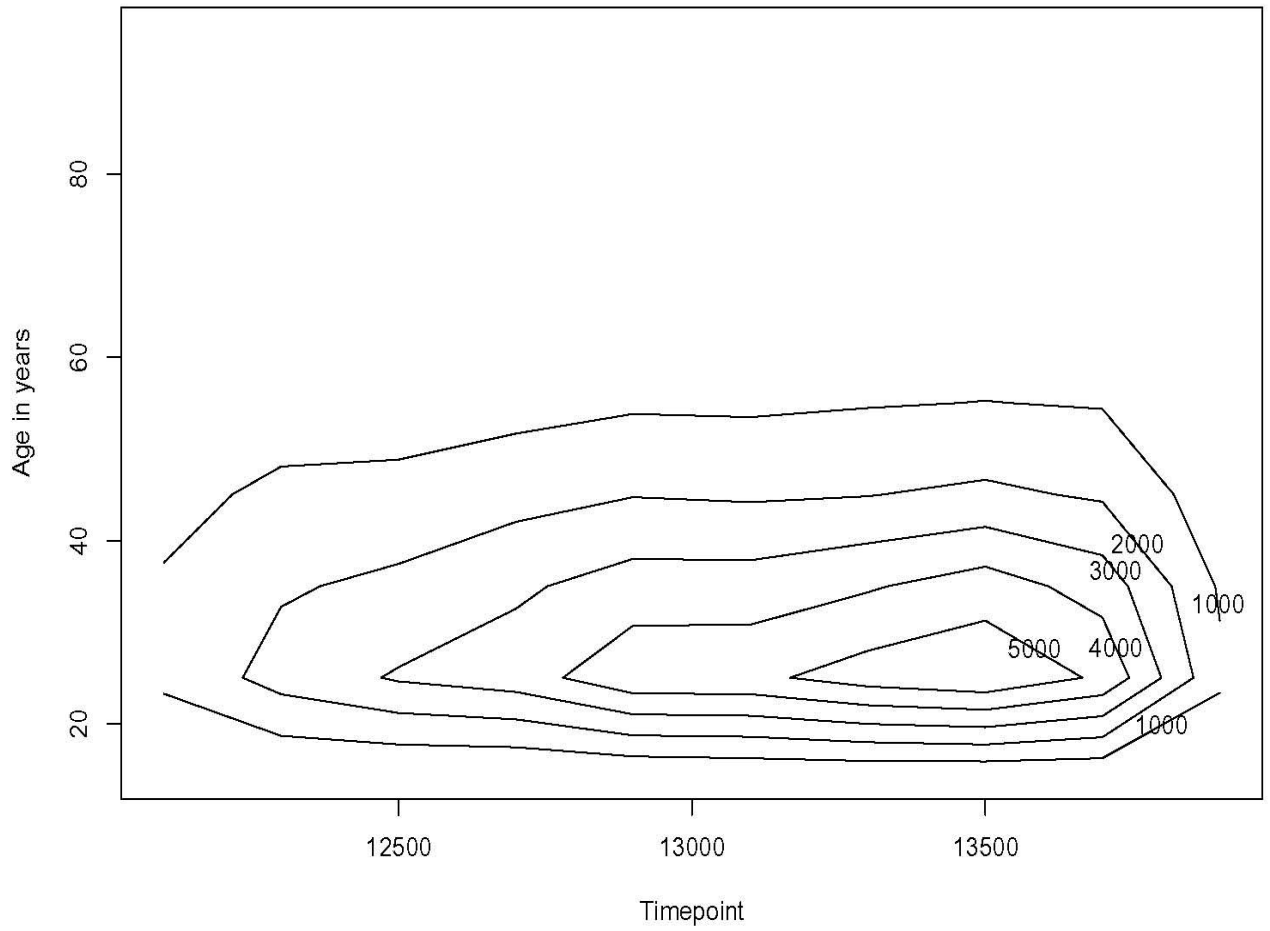
Problems with scatter plots



too much data $\Rightarrow$ over printing

Contour plot

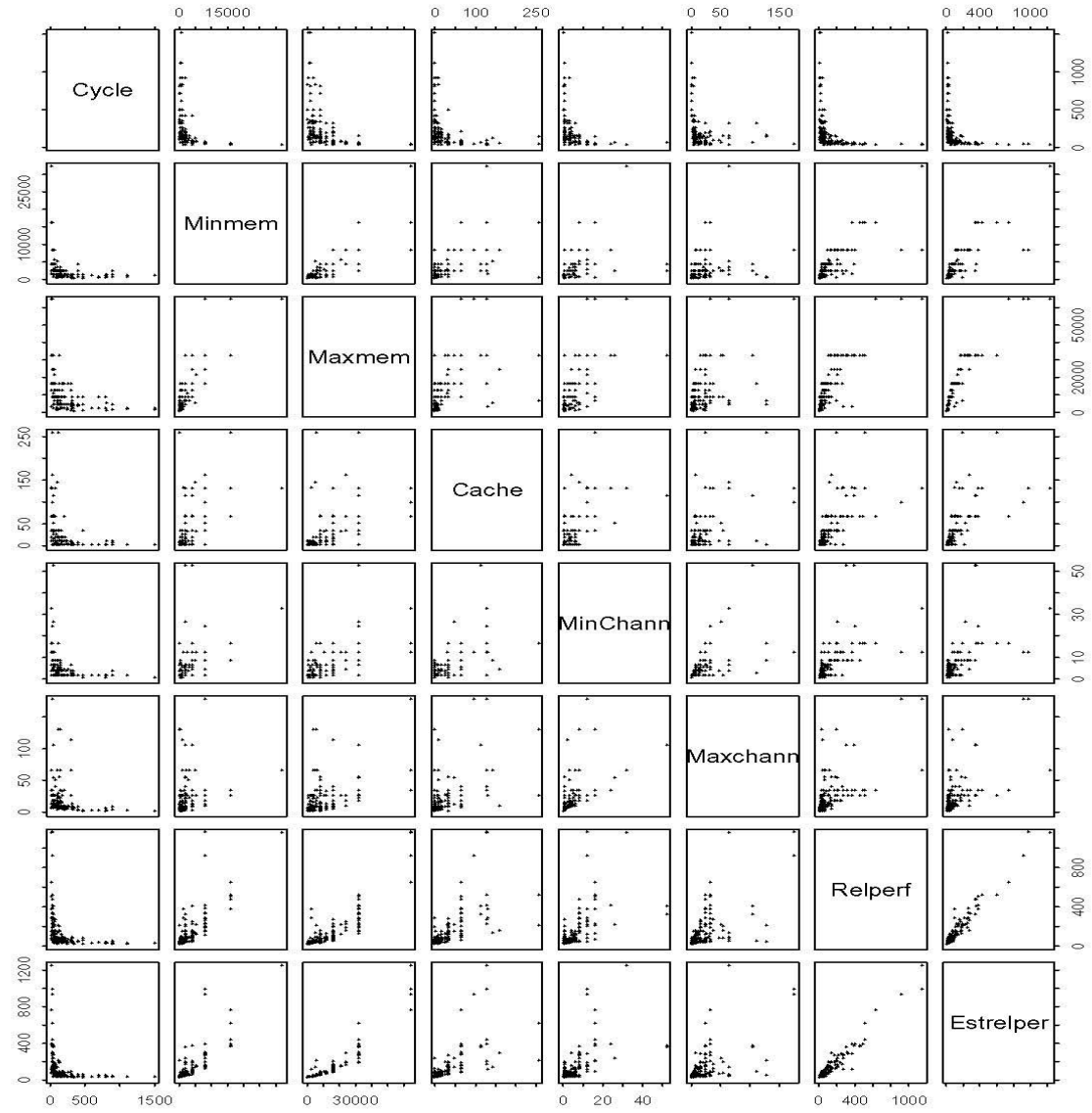
representing a 3-dimensional surface by plotting constant *z* slices, called contours, on a 2-dimensional format.



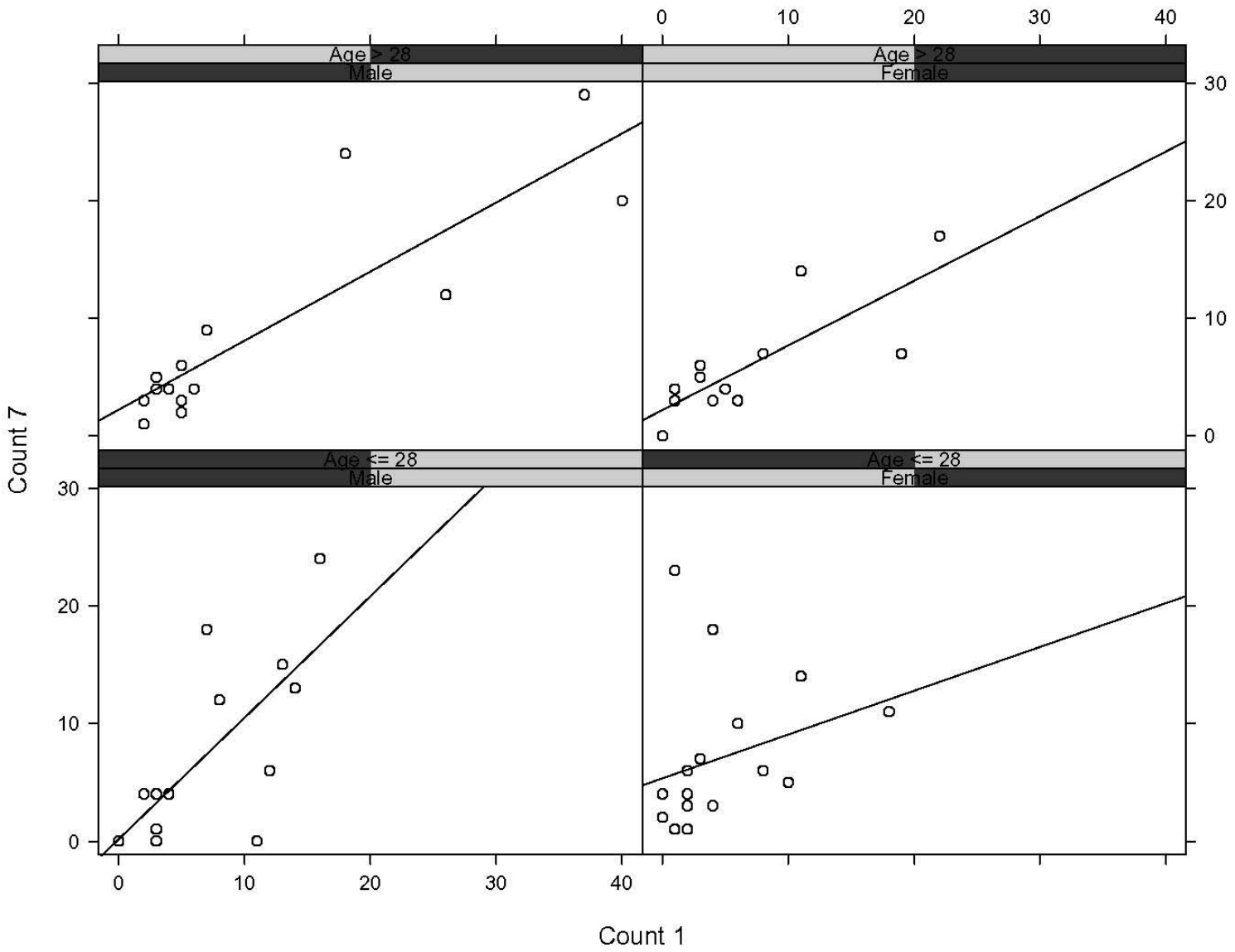
Displaying high-dimensional data

- multiple bivariate graphs
  - scatter plot matrix
  - trellis plot
- Icon plots
  - star graph
  - Chernoff's faces
- Parallel coordinates

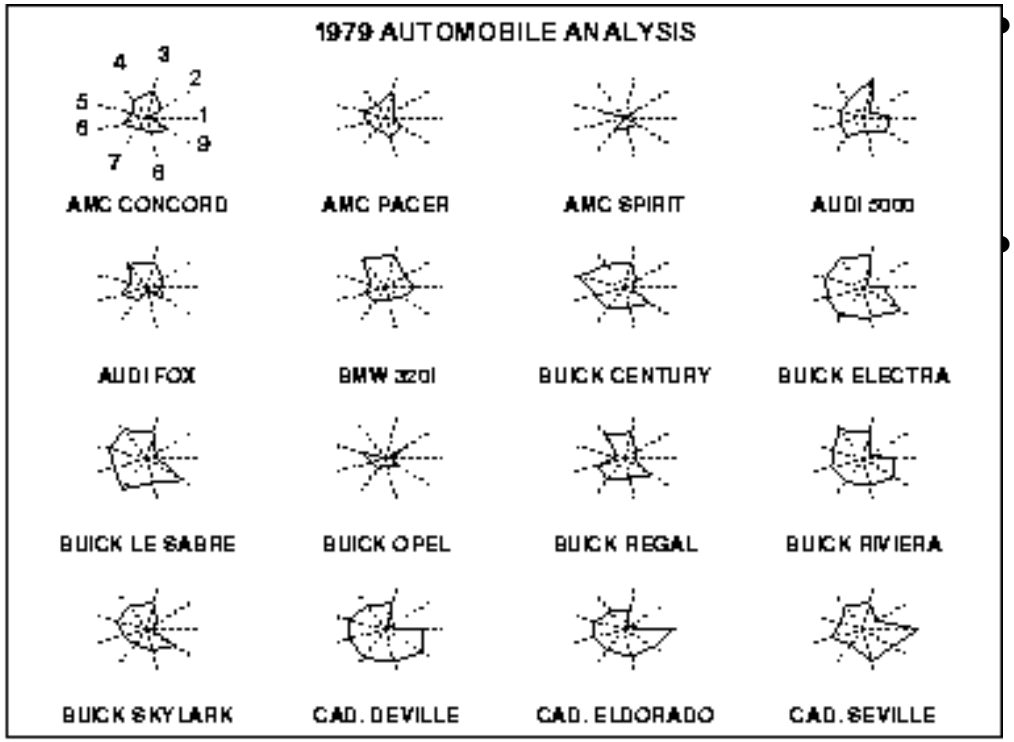
Scatter Plot Matrix



Trellis Plot



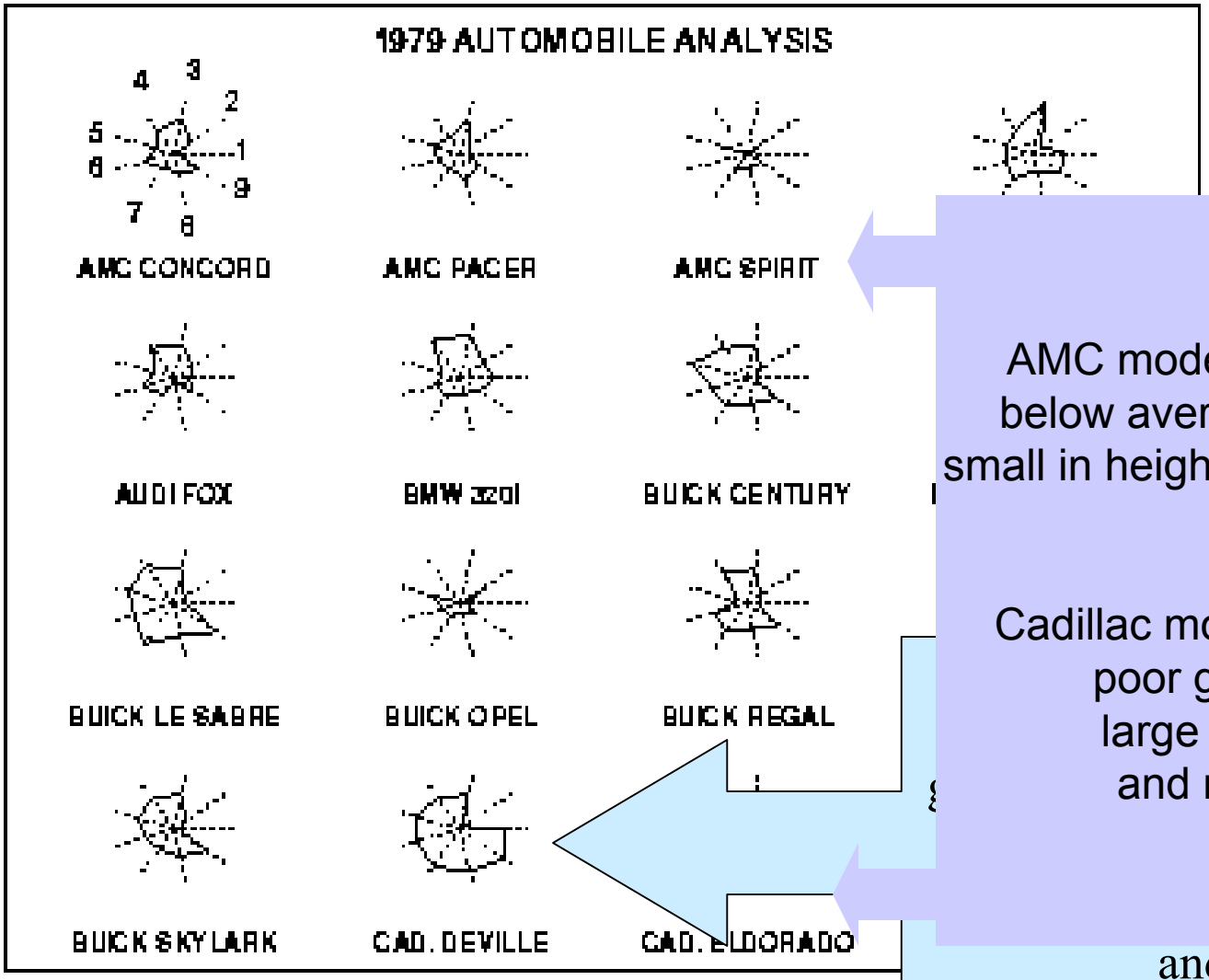
Star plots



1 Price
2 Mileage (MPG)
3 1978 Repair Record (1 = Worst, 5 = Best)
4 1977 Repair Record (1 = Worst, 5 = Best)
5 Headroom
6 Rear Seat Room
7 Trunk Space
8 Weight
9 Length

- Each star represents a single observation. Star plots are used to examine the relative values for a single data point
- The star plot consists of a sequence of equi-angular spokes, called radii, with each spoke representing one of the variables. The data length of a spoke is proportional to the magnitude of the variable for the data point relative to the maximum magnitude of the variable across all data points. A line is drawn connecting the data values for each spoke. This gives the plot a star-like appearance and the origin of the name of this plot.

Star plots cont.



1 Price
2 Mileage (MPG)
3 1978 Repair
4 1977 Repair

om

AMC models: inexpensive, below average gas mileage, small in height, weight roominess.

Cadillac models: expensive, poor gas mileage, large in both size and roominess.

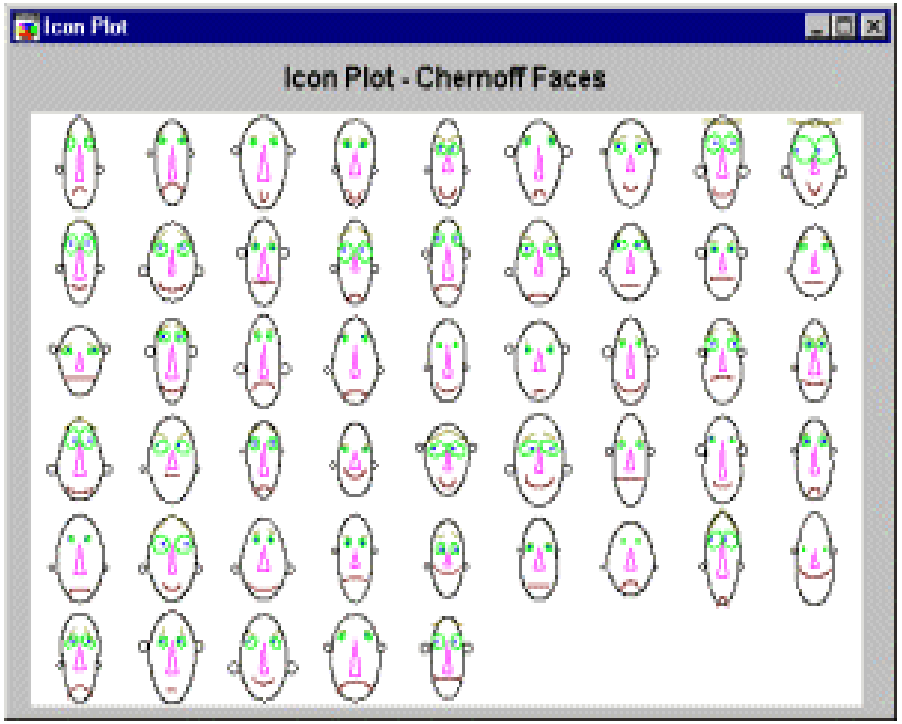
and size.

Chernoff's Faces

- described by ten facial characteristic parameters: head eccentricity, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, eye spacing, eye size, mouth length and degree of mouth opening

- Chernoff faces applet

- more icon plots



Dimension reduction

- scatter plot a simple example of projecting onto two dimensions

- Principle Component Analysis (PCA)
- Factor Analysis
- Multidimensional Scaling (MDS)

PCA

- Reference, http://www.cis.hut.fi/~jhollmen/dippa/node30.html

PCA

- linear transform widely used in data analysis and compression
- Intuition:
  - compute correlation matrix
  - compute a new set of artificial variables (eigenvectors), each with its associated eigenvalue
  - eigenvalue represents the contribution of each artificial variable to the total variation in the dataset
  - eigenvectors are the principle component vectors
  - new data vectors are formed by projecting the original data onto the principle component vectors

PCA cont.

- data vectors, x(i), i = 1, …, n
  Assume the estimated mean has been subtracted from each column
- correlation matrix:

$$R = E(xx^T) = \frac{1}{n}\sum_{i=1}^{n} x(i)x(i)^T$$

reference: Peter Andras, http://www.staff.ncl.ac.uk/peter.andras/pcaic

PCA

- Let X be p x n data matrix, each row is data vector x(i)
- Assume the estimated mean has been subtracted from each column

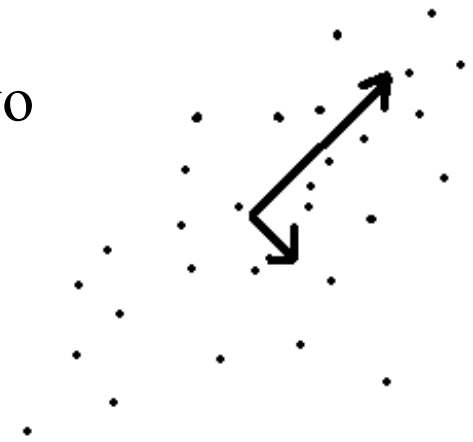
PCA cont.

- The eigenvectors are determined by the equation:

$$Rv = \lambda v$$

where λ is a real number.

Example with two eigenvectors:



PCA cont.

- In principle we should find p eigenvectors if the dimensionality of the data vectors is p.
- If the data vectors are situated on a lower dimensional linear surface we find less than p eigenvectors (i.e., the determinant of the correlation matrix is zero).
- If v^1, v^2, …, v^m, m<p, are the eigenvectors of R then the new, transformed data vectors are calculated as:

$$y(i) = (y_1(i), \ldots, y_m(i))$$

$$y_k(i) = x(i)^T v^k$$

- What is the complexity of computing these eigenvectors for R?

PCA example

- The weights (loadings) defining the components may provide insight into the 'artificial' variables
- Huba, et al. (1981) example:
  - data on 1684 students in LA showing consumption of 13 legal and illegal substances:
    cigarettes, beer, wine, spirits, cocaine, tranquilizers, drug store medications, heroin, marijuana, hashish, inhalants, hallucinogenics and amphetamines
  - scored: 1 – never tried, 2 - tried only once, 3 – tried a few times, 4 – tried many times, 5 – used regularly
  - first principle component:
    (0.278,0.286,0.265,0.318,0.208,0.293,0.176,0.202,0.339,0.329,0.276,0.248,0.329)
  - *interpretation: biggest difference among students is in terms of how often they use substances, regardless of which substance they use*
  - second principle component:
    (0.280,0.396,0.392,0.325,-0.288,-0.259,-0.189,-0.315,0.163,-0.50,-0.169,-0.329,-0.232)
  - *interpretation: after controlling for overall substance use, the major differences is between students is based on whether they use legal or illegal substances*

Multidimensional Scaling (MDS)

- MDS attempt to represent data points in a lower dimensional space while preserving, as much as possible the distances between the data points.
- Methods differ based on how they define the distances that are preserved, the distances they map to and how the calculations are performed.
- PCA is the simplest form: distances are Euclidean, they are mapped to distances in a reduced space that are also measured using Euclidean distance. The sum of squared distances between the original data points and their projection defines the quality of the representation.
- With MDS one may analyze any kind of similarity or dissimilarity matrix, in addition to correlation matrices.

- The following simple example may demonstrate the logic of an MDS analysis. Suppose we take a matrix of distances between major US cities from a map. We then analyze this matrix, specifying that we want to reproduce the distances based on two dimensions. As a result of the MDS analysis, we would most likely obtain a two-dimensional representation of the locations of the cities, that is, we would basically obtain a two-dimensional map.
- In general then, MDS attempts to arrange "objects" (major cities in this example) in a space with a particular number of dimensions (two-dimensional in this example) so as to reproduce the observed distances. As a result, we can "explain" the distances in terms of underlying dimensions; in our example, we could explain the distances in terms of the two geographical dimensions: north/south and east/west.
- **Orientation of axes.** As in factor analysis, the actual orientation of axes in the final solution is arbitrary. To return to our example, we could rotate the map in any way we want, the distances between cities remain the same. Thus, the final orientation of axes in the plane or space is mostly the result of a

- MDS is not so much an exact procedure as rather a way to "rearrange" objects in an efficient manner, so as to arrive at a configuration that best approximates the observed distances. It actually moves objects around in the space defined by the requested number of dimensions, and checks how well the distances between objects can be reproduced by the new configuration. In more technical terms, it uses a function minimization algorithm that evaluates different configurations with the goal of maximizing the goodness-of-fit (or minimizing "lack of fit").
- **Measures of goodness-of-fit: Stress.** The most common measure that is used to evaluate how well (or poorly) a particular configuration reproduces the observed distance matrix is the stress measure. The raw stress value *Phi* of a configuration is defined by:
- $\text{Phi} = [d_{ij} - f(\ _{ij})]^2$
- In this formula, d_{ij} stands for the reproduced distances, given the respective number of dimensions, and $_{ij}$ ($delta_{ij}$) stands for the input data (i.e., observed distances). The expression $f(\ _{ij})$ indicates a *nonmetric*, monotone transformation of the observed

Next Time

- Reading:
  - HMS, chapter 3
  - Shneiderman article

- Topic:
  - Guest lecture by Ben Shneiderman on Information Visualization! Don't miss this one!

References

- *Principles of Data Mining*, Hand, Mannila, Smyth. MIT Press, 2001.
- NIST/Semantech Engineering Statistics Handbook, http://www.itl.nist.gov/div898/handbook/eda/eda.htm
- Stat soft Electronic Textbook http://www.statsoftinc.com/textbook/stathome.html
- Peter Andras, http://www.staff.ncl.ac.uk/peter.andras/pcaica.ppt