Exploring data

What is data exploration?

- To better understand the characteristics of data
  - Help to select the right tool for preprocessing or analysis
  - Help to recognize patterns (사람의 직관을 통한 데이터 패턴 인식)

- Data exploration
  - Summary statistics
  - Visualization
  - Multidimensional data analysis

Sample datasets

- Can be obtained from the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/

Summary statistics

Summary Statistics

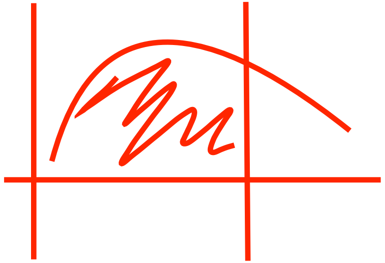
- Summary statistics are numbers that summarize properties of the data
  - Summarized properties include frequency, location and spread
  - Examples: location - mean
    spread - standard deviation

Frequency and Mode

- The frequency of an attribute value
  - The percentage of time the value occurs in the data set
  - Example: given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode of an attribute
  - the value that appears most frequently
- The notions of frequency and mode are typically used with categorical data
- Example: Table 3.1

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 $x_{50\%}$.



x 75%

Measures of Location: Mean and Median

평균 중간값

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

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

- Mean
  - The most common measure of the location of a set of points
  - Very sensitive to outliers
  - The median or a trimmed mean is also commonly used.

Measures of Spread: Range and Variance

- Range: the difference between the max and min
- The variance or standard deviation: the most common measure of the spread of a set of points.

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

- Sensitive to outliers, so absolute average deviation (AAD), median absolute deviation (MAD), and interquartile range (IQR) are often used.

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

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

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

Visualization

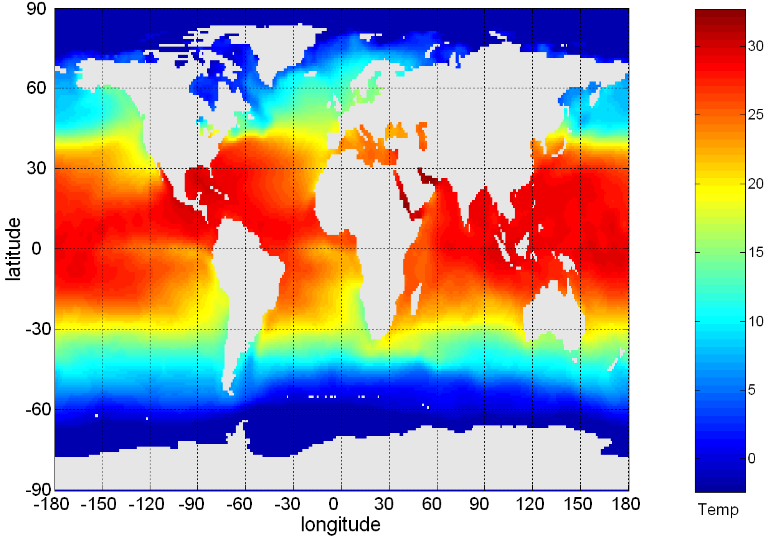
Visualization

- The conversion of data into a visual or tabular format
- The characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
- 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

Example: Sea Surface Temperature

The following shows the Sea Surface Temperature (SST) for July 1982

- Tens of thousands of data points are summarized in a single figure



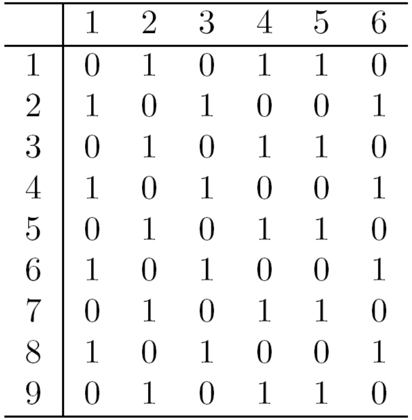
Representation

- Map 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.

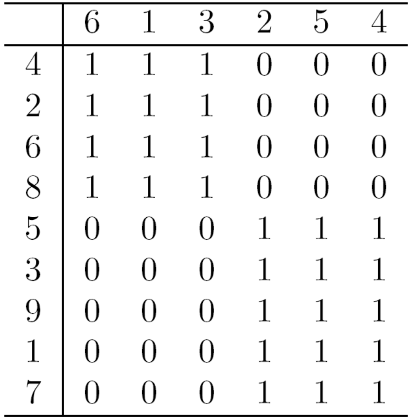
관계를 나타냄

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

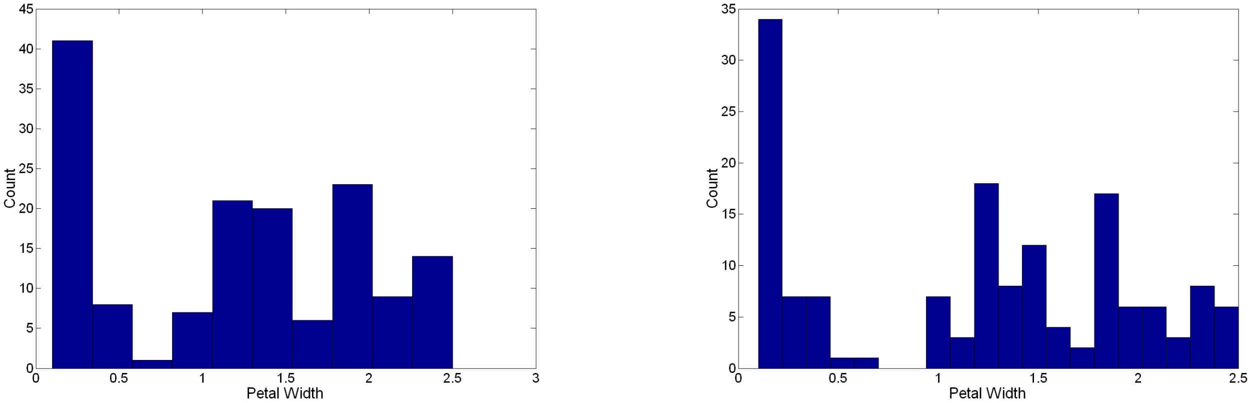
Selection

- The elimination or the de-emphasis of certain objects and attributes
- 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
- 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

Visualization Techniques: Histograms

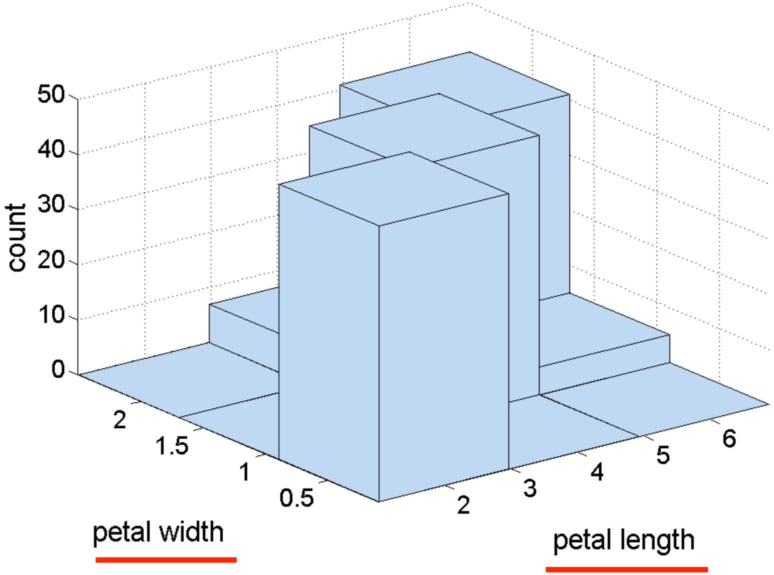
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)

간격



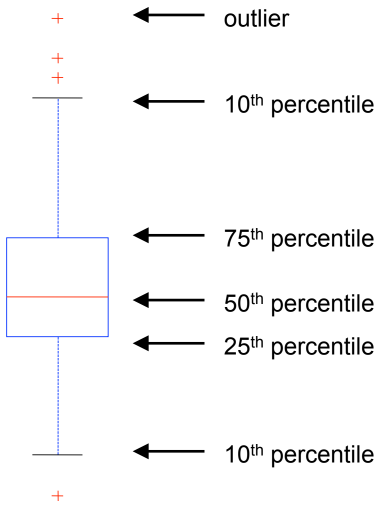
Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



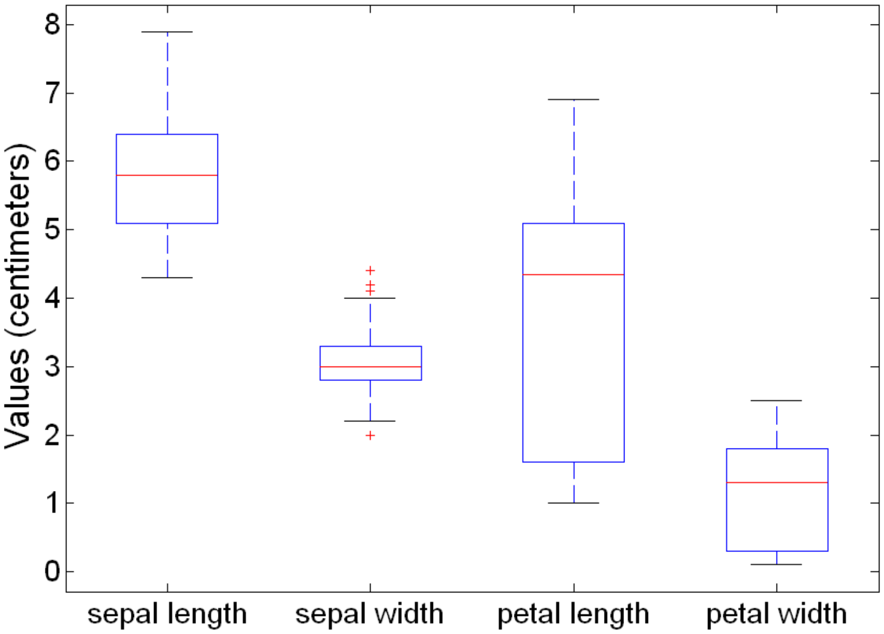
Visualization Techniques: Box Plots

- Box Plots
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



Example of Box Plots

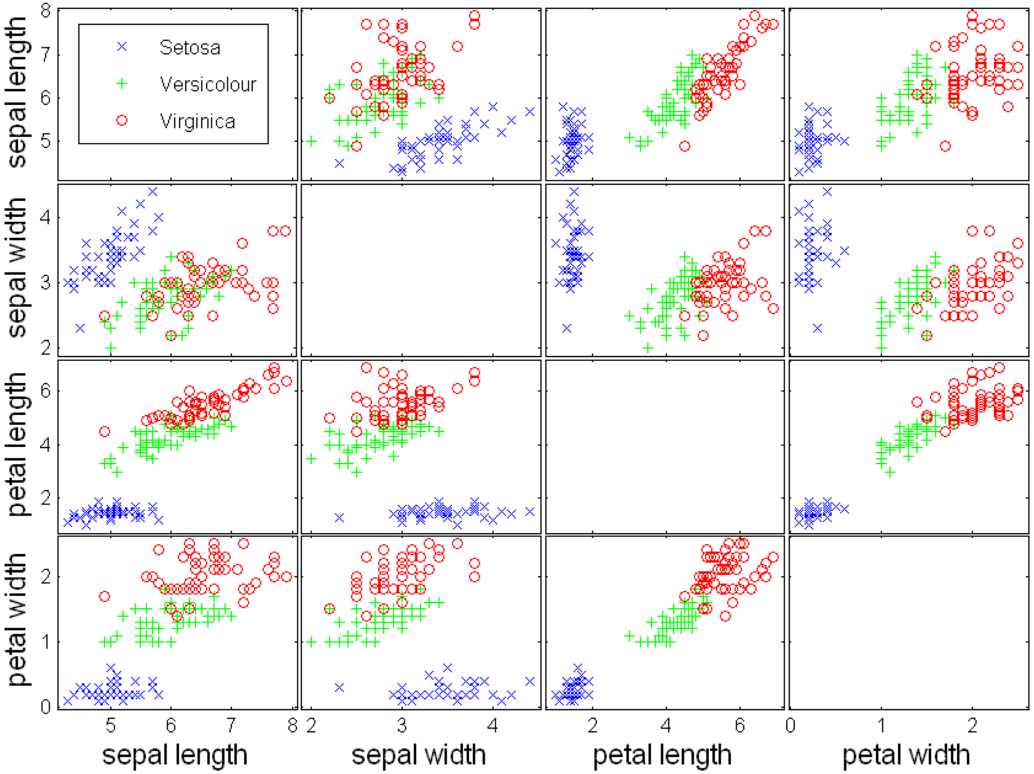
- Box plots can be used to compare attributes



Visualization Techniques: Scatter Plots

- Scatter plots
  - Attributes’ values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes

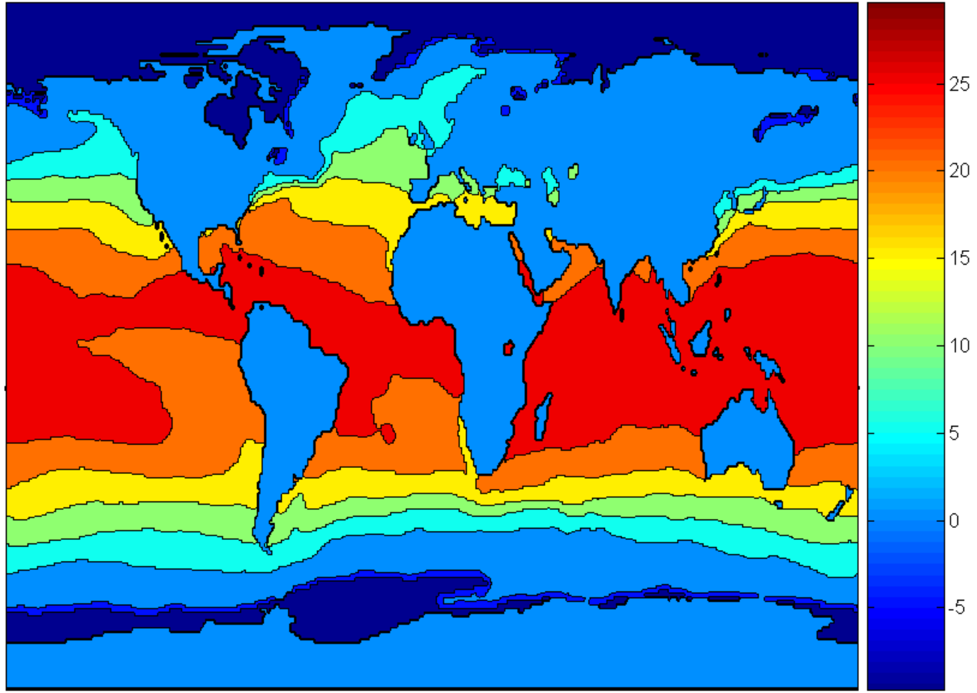
Scatter Plot – example



Visualization Techniques: Contour Plots

- 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.

Contour Plot Example

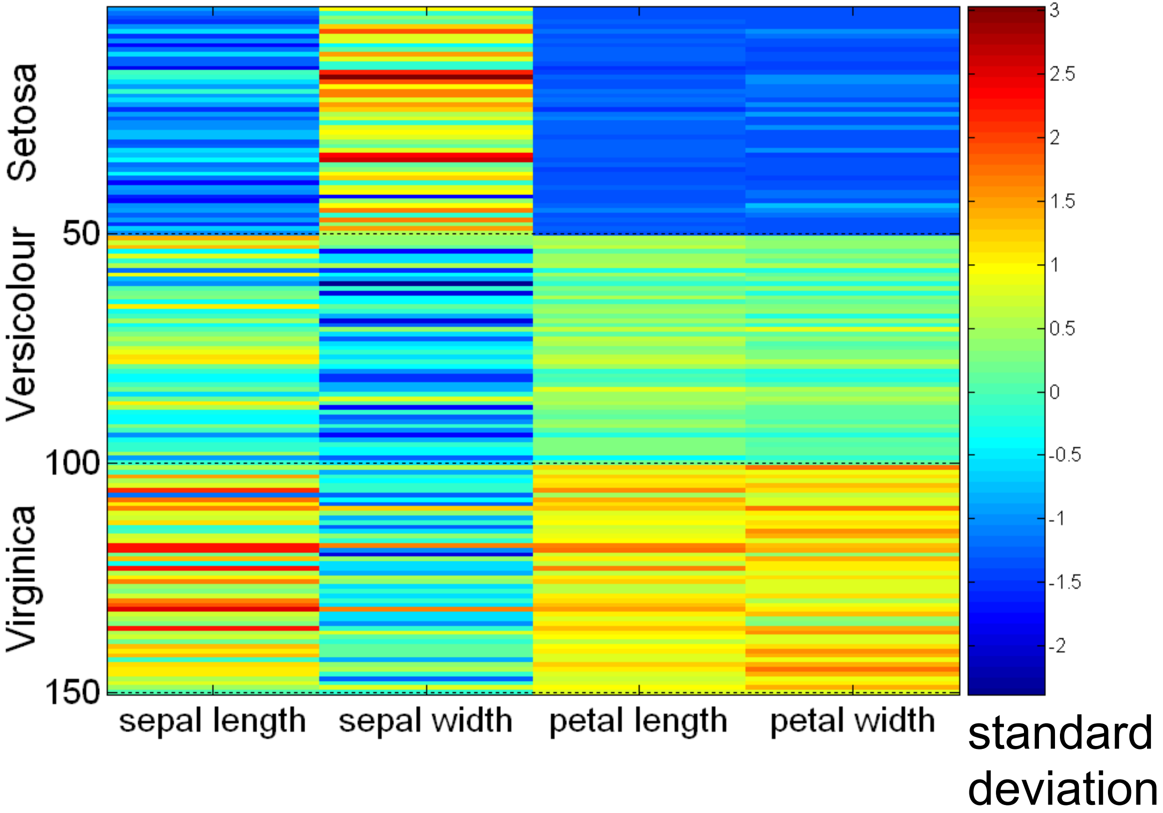


Celsius

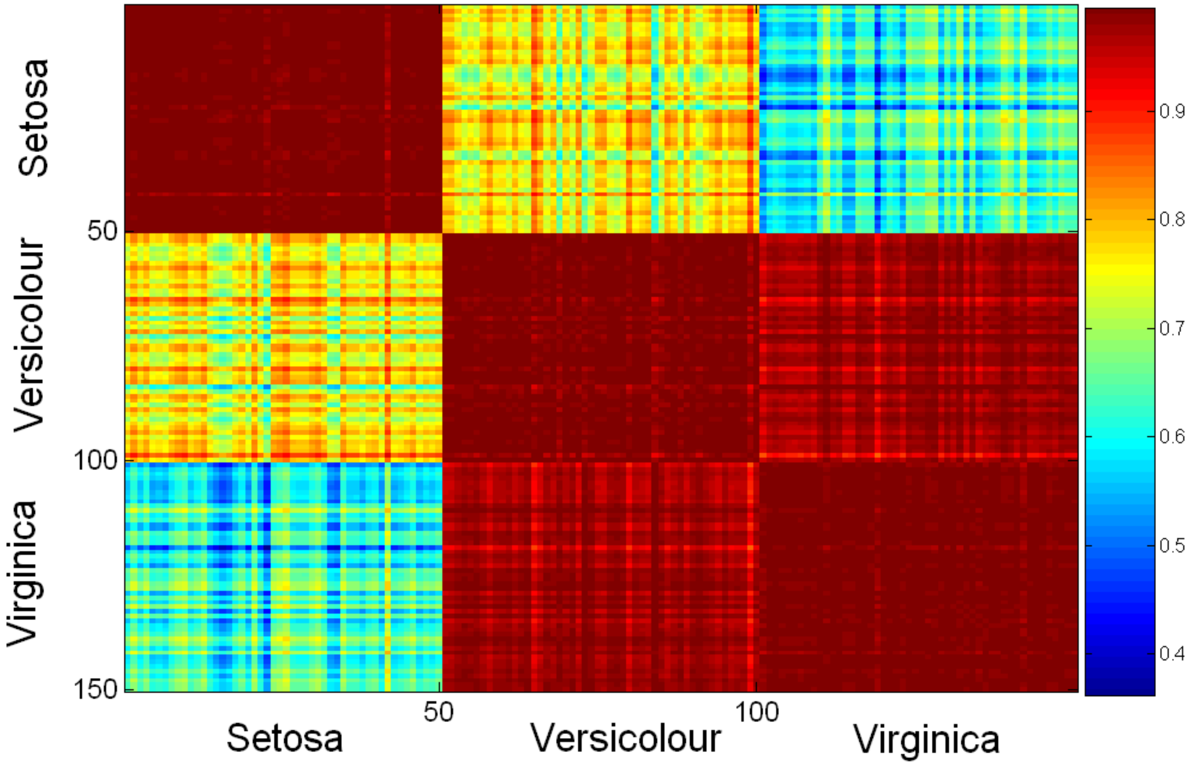
Visualization Techniques: Matrix Plots

- Matrix plots
  - Can plot the data matrix
  - This can be useful when objects are sorted according to class
  - Typically, the attributes are normalized to prevent one attribute from dominating the plot
  - Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects
  - Examples of matrix plots are presented on the next two slides

Matrix Plot – example (data)



Matrix Plot – example (correlation)

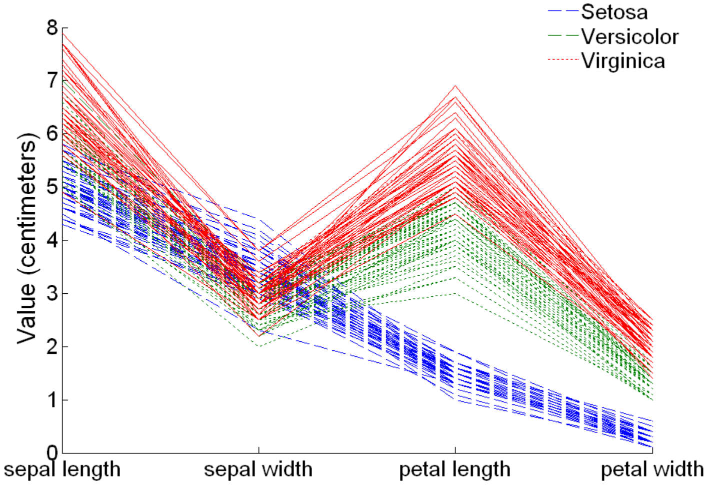


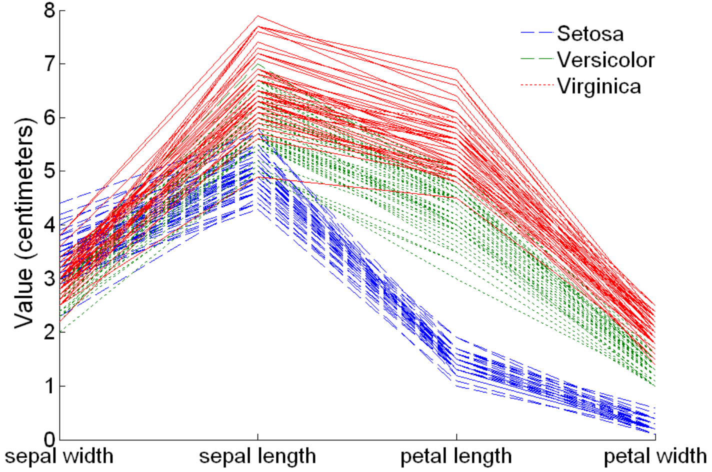
Visualization Techniques: Parallel Coordinates

- Parallel Coordinates
  - Used to plot the attribute values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line

  - Each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some attributes
  - Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots – example

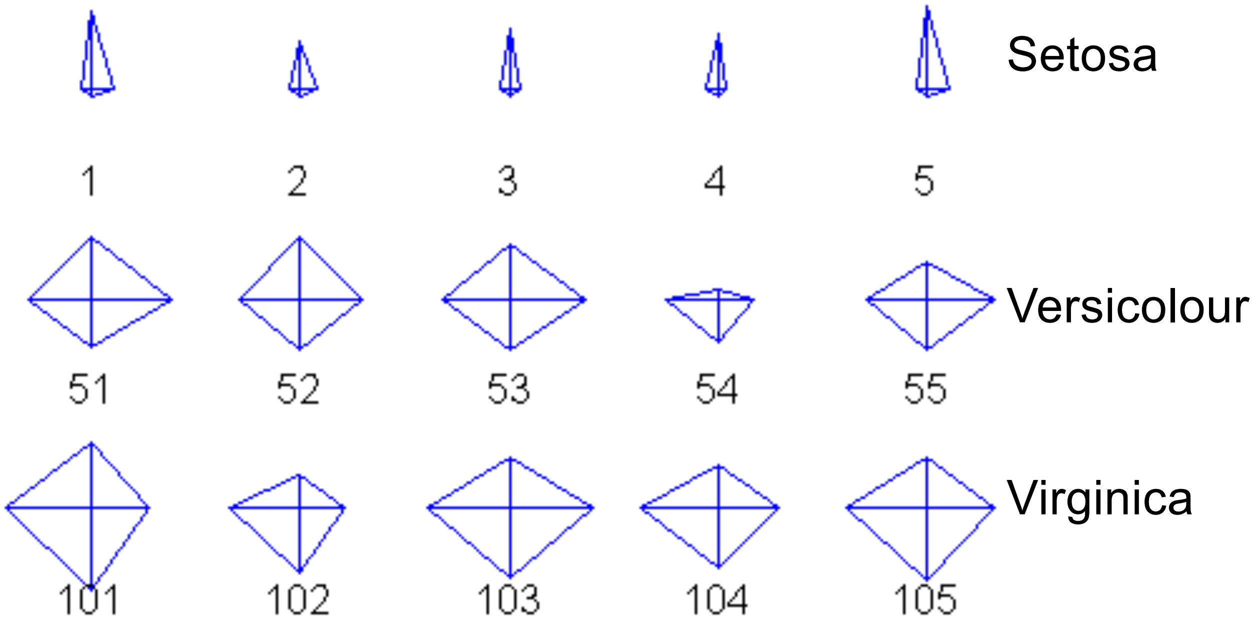




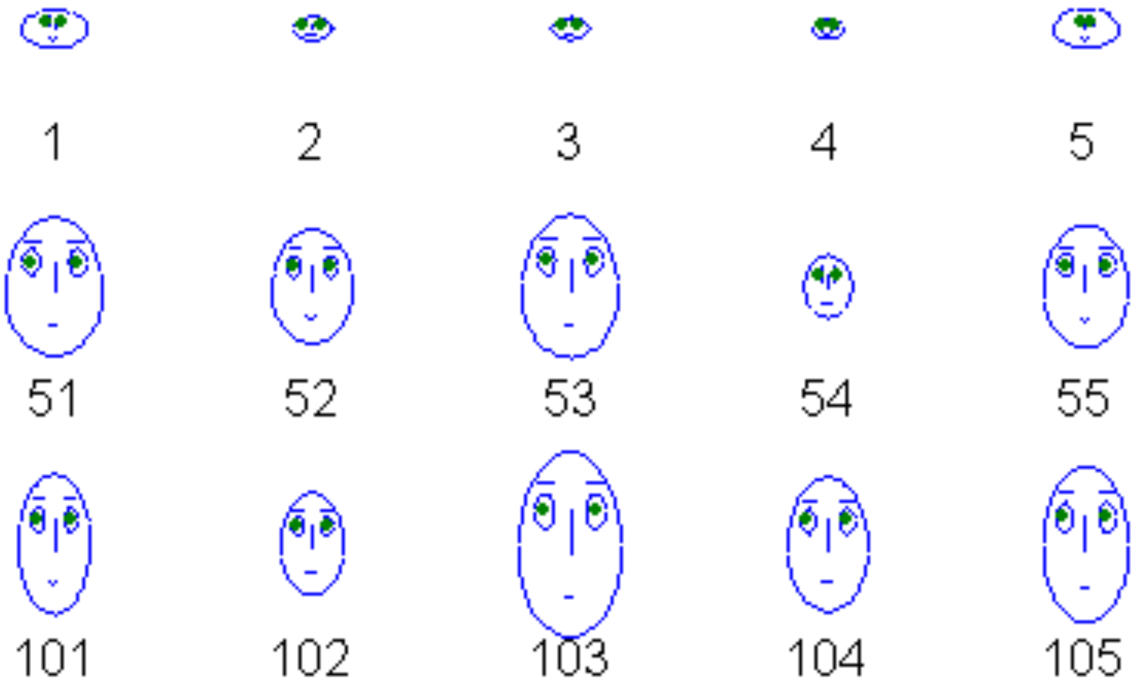
Other Visualization Techniques

- Star Plots
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon
- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human's ability to distinguish faces

Star Plots


1
2
3
4
5
Setosa
51
52
53
54
55
Versicolour
101
102
103
104
105
Virginica

Chernoff Faces



Setosa

Versicolour

Virginica