EXPLORING DATA

Data Exploration

- A preliminary exploration of the data to better understand its characteristics.
- Helping to select the right tool for preprocessing or analysis
- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization

Iris Sample Data Set

Many of the exploratory data techniques are illustrated with the Iris Plant data set.

Can be obtained from the UCI Machine Learning Repository
http://www.ics.uci.edu/~mlearn/MLRepository.html

- Three flower types (classes):
  - Setosa
  - Virginica
  - Versicolour
- Four (non-class) attributes
  - Sepal width and length
  - Petal width and length

Summary Statistics

Summary statistics are quantities, such as the mean and standard deviation, that capture various characteristics of a potentially large set of values with a single number or a small set of numbers.

Summarized properties include frequency, location and spread

Examples: location - mean

spread - standard deviation

Frequencies and the 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 nominal data

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 $x_{50\%}$

$$\min(x) = x_{0\%} \text{ and } \max(x) = x_{100\%}$$

Example

Percentile	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	2.0	1.0	0.1
10	4.8	2.5	1.4	0.2
20	5.0	2.7	1.5	0.2
30	5.2	2.8	1.7	0.4
40	5.6	3.0	3.9	1.2
50	5.8	3.0	4.4	1.3
60	6.1	3.1	4.6	1.5
70	6.3	3.2	5.0	1.8
80	6.6	3.4	5.4	1.9
90	6.9	3.6	5.8	2.2
100	7.9	4.4	6.9	2.5

Measures of Location: Mean and Median

Mean

mean is the most common measure of the location of a set of points.

$\{x_1, \ldots, x_m\}$

attribute values of *x* for *m* objects.

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

mean is very sensitive to outliers

Median

$\{x_{(1)}, \ldots, x_{(m)}\}$

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

trimmed mean

trimmed mean

- ✓ A percentage *p* between 0 and 100 is specified
- ✓ top and bottom (*p*/2)% of the data is thrown out
- ✓ mean is then calculated in the normal way.

Measure	Sepal Length	Sepal Width	Petal Length	Petal Width
mean	5.84	3.05	3.76	1.20
median	5.80	3.00	4.35	1.30
trimmed mean (20%)	5.79	3.02	3.72	1.12

Measures of Spread: Range and Variance

- ✓ Range is the difference between the max and min
- ✓ Variance or standard deviation is 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

absolute average deviation (AAD)

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

median absolute deviation (MAD)

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

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

Multivariate Summary Statistics

if x_i and x_j are the i_{th} and j_{th} attributes, then

$$s_{ij} = \text{covariance}(x_i, x_j)$$

$$\text{covariance}(x_i, x_j) = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})$$

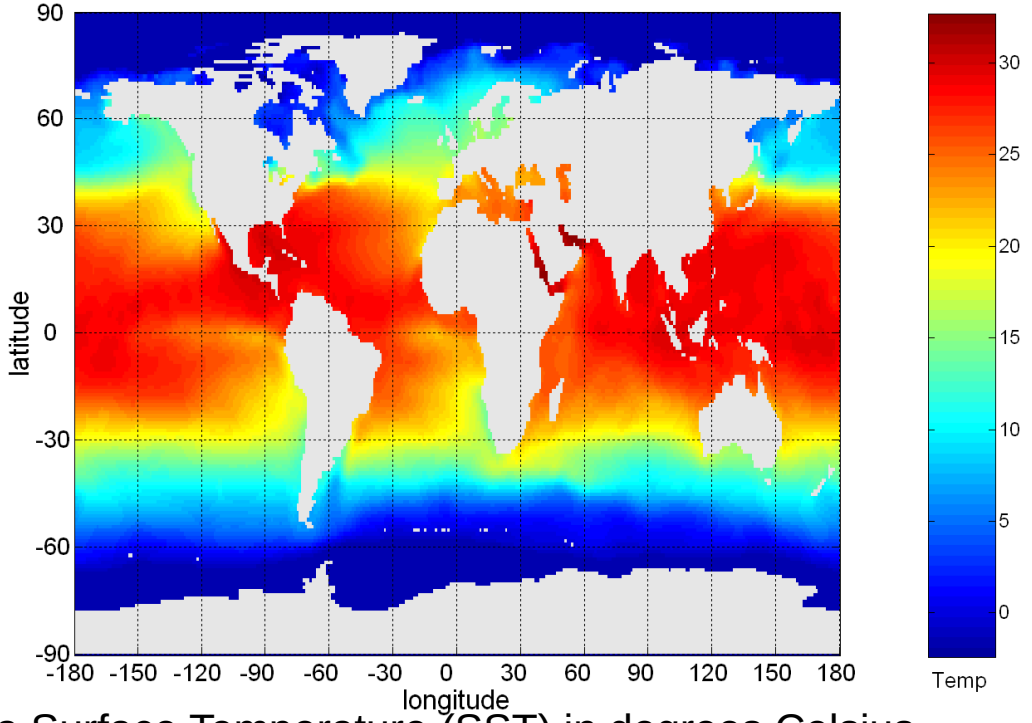
$$r_{ij} = \text{correlation}(x_i, x_j) = \frac{\text{covariance}(x_i, x_j)}{s_i s_j}$$

correlation matrix R

Visualization

Visualization

- ✓ Visualization is the conversion of data into a visual format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported
- ✓ people can quickly absorb large amounts of visual information



Sea Surface Temperature (SST) in degrees Celsius

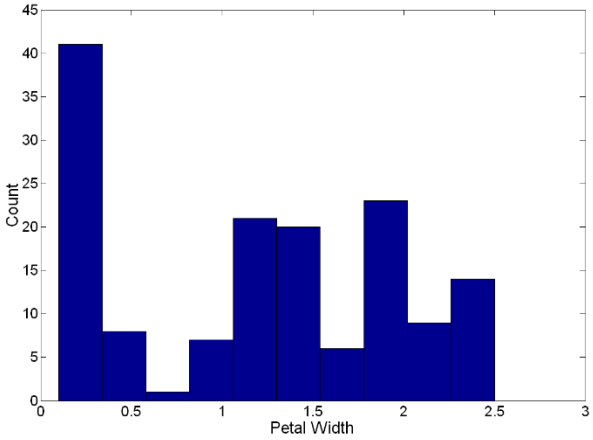
Visualization Techniques: Histograms

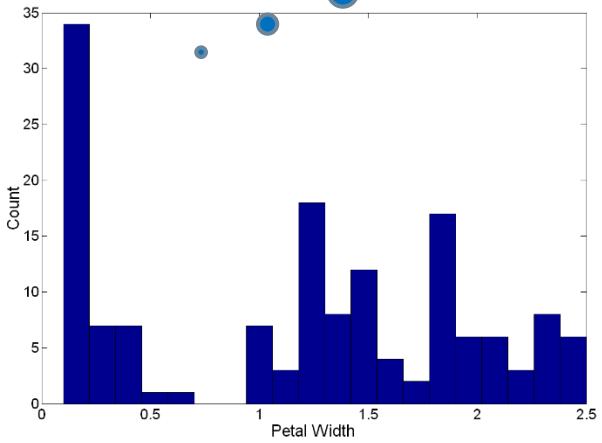
Histogram

- 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)

relative frequency histogram

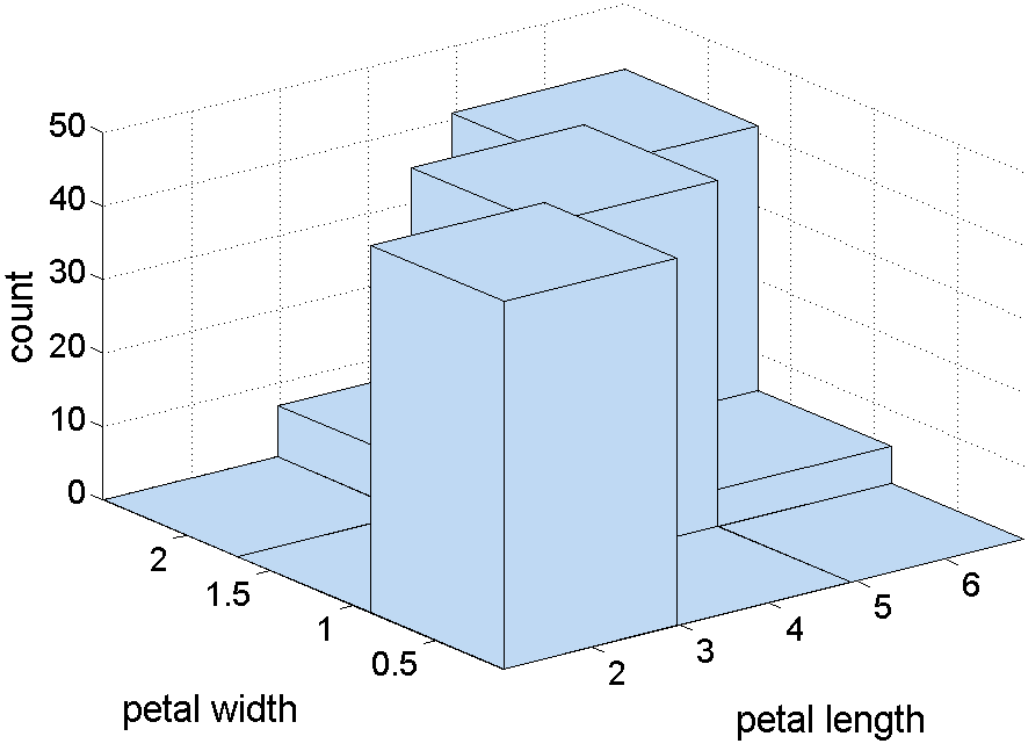




Two-Dimensional Histograms

Show the joint distribution of the values of two attributes

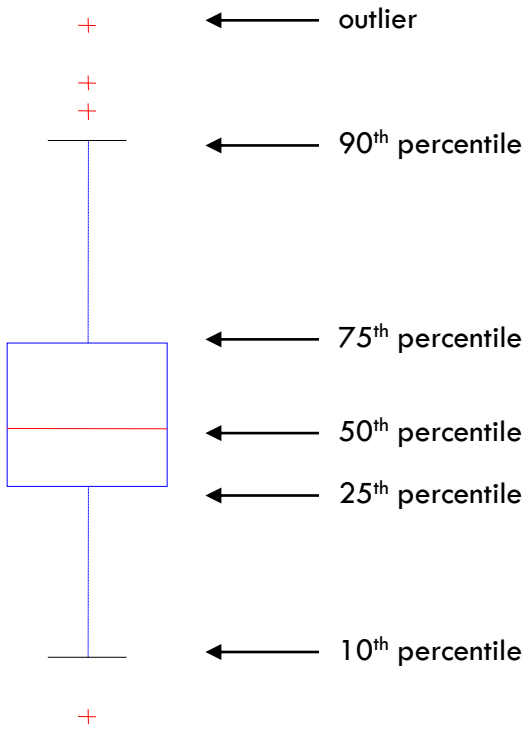
Example: petal width and petal length



Visualization Techniques: Box Plots

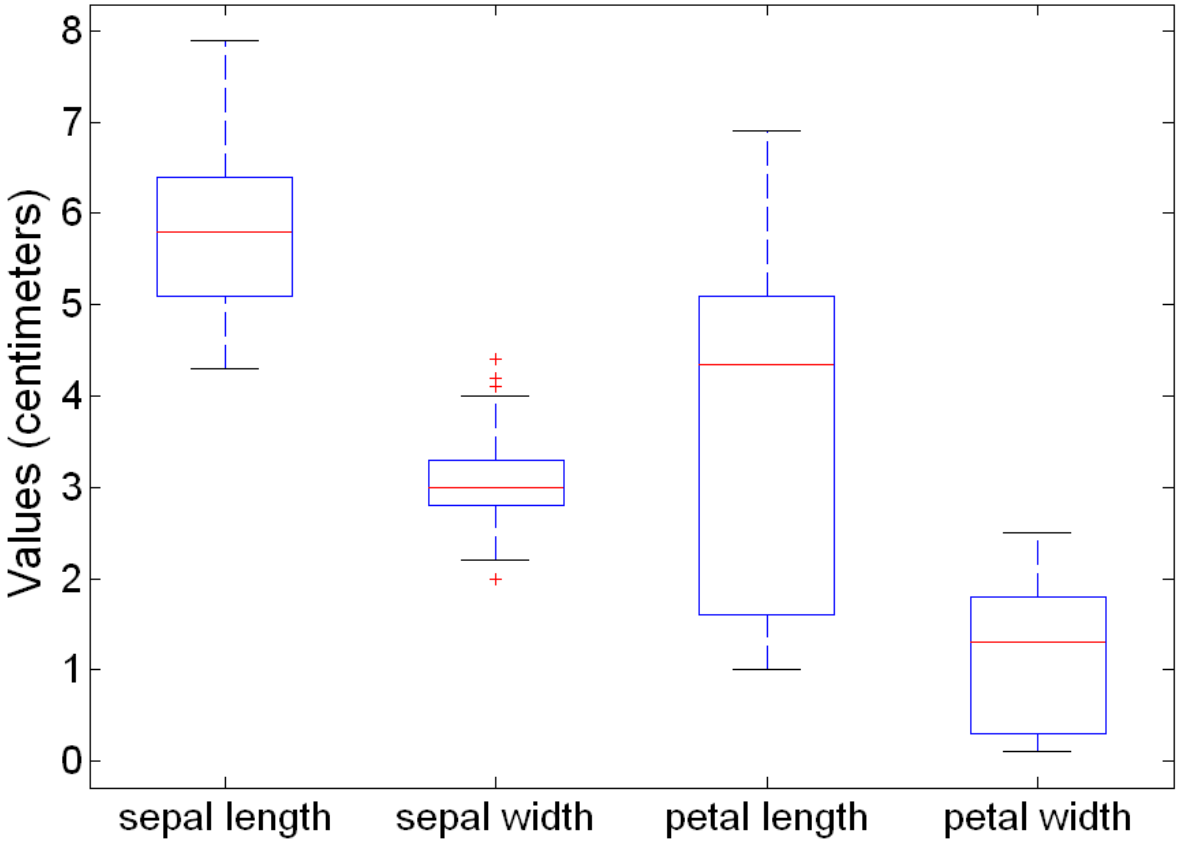
Box Plots
Another way of displaying the distribution of data

Following figure shows the basic part of a box plot



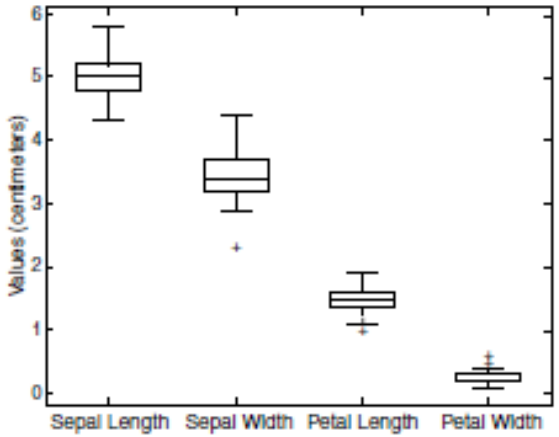
Visualization Techniques: Box Plots

Box plots can be used to compare attributes

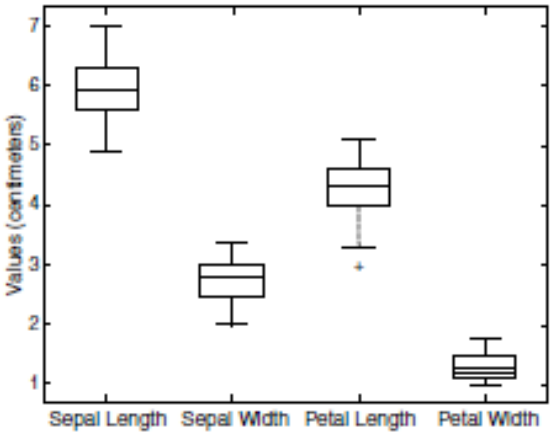


Visualization Techniques: Box Plots

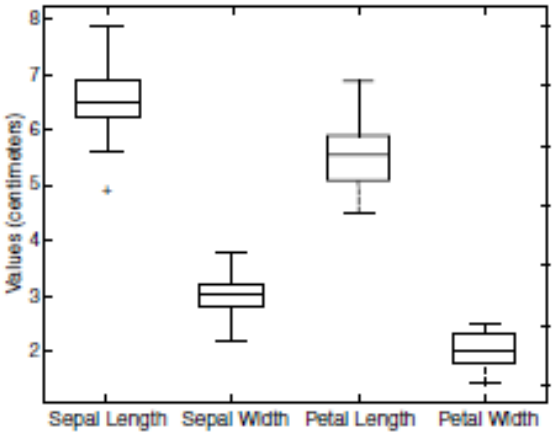
Box plots can also be used to compare how attributes vary between different classes of objects



(a) Setosa.



(b) Versicolour.



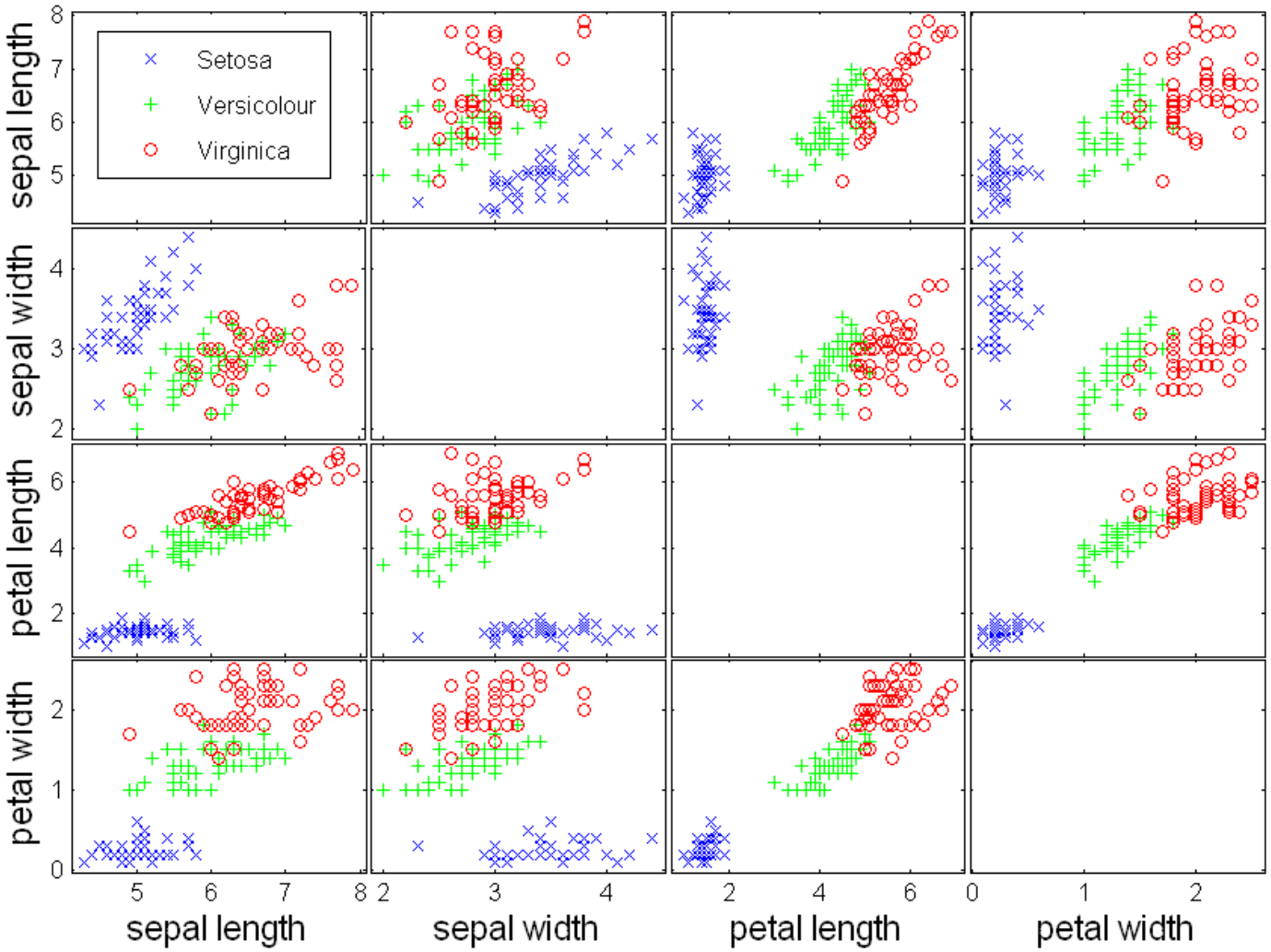
(c) Virginica.

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

Visualization Techniques: Scatter Plots

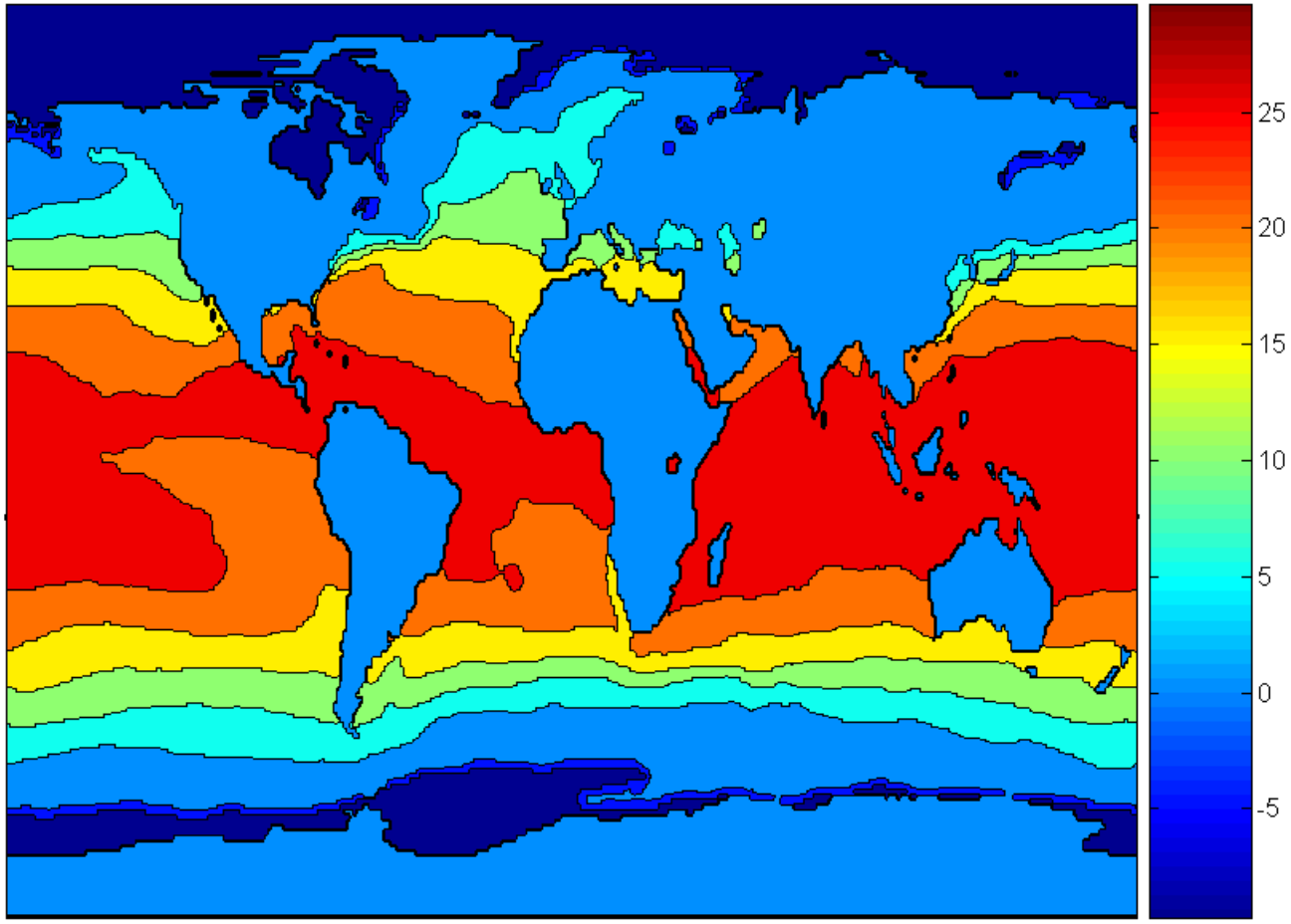


Visualization Techniques: Contour Plots

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.

Visualization Techniques: Contour Plots

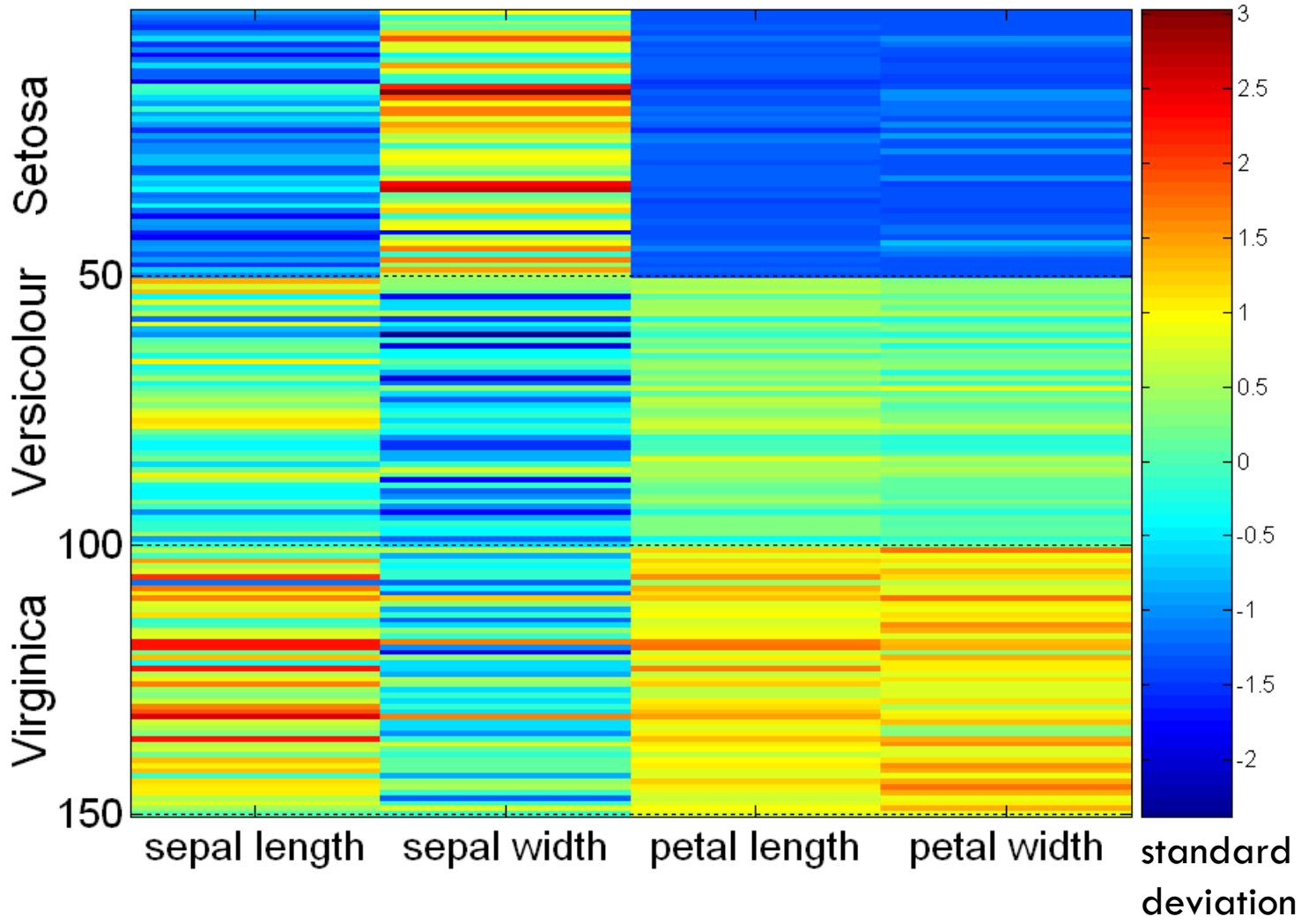


Visualization Techniques: Matrix Plots

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

Visualization Techniques: Matrix Plots



Visualization Techniques: Matrix Plots