Visualization

Dr. Sanjay Ranka
Professor
Computer and Information Science and Engineering
University of Florida

What is visualization ?

- Visualization is the process of converting data (information) in to a visual format to analyze
  - the characteristics of the data
  - the relationships among data items or attributes
- E.g. graphs and tables

What is visualization ? …

- Two steps:
  - Mapping the objects, attributes and relationships involved in a set of data on to visual objects, attributes and relationships
  - Interpretation of the visualized information by the observing human being and the formation of a mental model of the information

Motivation for Visualization

- Humans are good at quickly absorbing large amounts of visual information and finding patterns in such information
- To make use of the domain knowledge that is “locked up in people’s heads”
- Useful as a supplementary technique as a part of data exploration in order to get an idea of the nature of the data

Visualizing Different Types of Data

- Type of data greatly affects how it can be plotted for visualization
- Dimensionality has a very strong effect. We can easily visualize one, two or three dimensional data, but its not very clear how multi-dimensional data should be visually represented
- Continuous vs. categorical attributes
- Nominal vs. ordinal attributes

General Categories of Visualization

- One way to classify categories of visualization is based on type of data:
  - One dimensional
  - Two dimensional
  - Three dimensional
  - Multi-dimensional
  - Hierarchical
  - Graph

General Categories of Visualization …

- Another way is to classify based on generic classes of applications:
  - **Scientific visualization:** The basic data concerns physical objects e.g. sea temperature data
  - **Information visualization:** The data is physical but rather is more conceptual or symbolic e.g. set of documents
  - **Statistical graphics:** Typical multivariate continuous or categorical data of type traditionally associated with statistics e.g. results of a survey

Representation

- Mapping data to graphical elements
- First step in visualization
- Three general approaches:
  - If only a single categorical variable of the object is being considered, then objects are often lumped in to categories based on the value of that attribute. These categories are then displayed as an entry in a table or bar charts
  - If an object has multiple attributes, it is displayed as a row (or column) in a table or as a line on a graph
  - Sometimes the object can be represented as a point in two-dimensional or three-dimensional space

Representation …

- For single attributes, the representation depends on the type of attribute
  - Ordinal and Continuous attributes can be mapped to continuous, order graphical features: location along x, y, z axes, intensity, color or size
  - Categorical attributes can be mapped to distinct graphical features: position, color, shape, orientation, etc.

Representation …

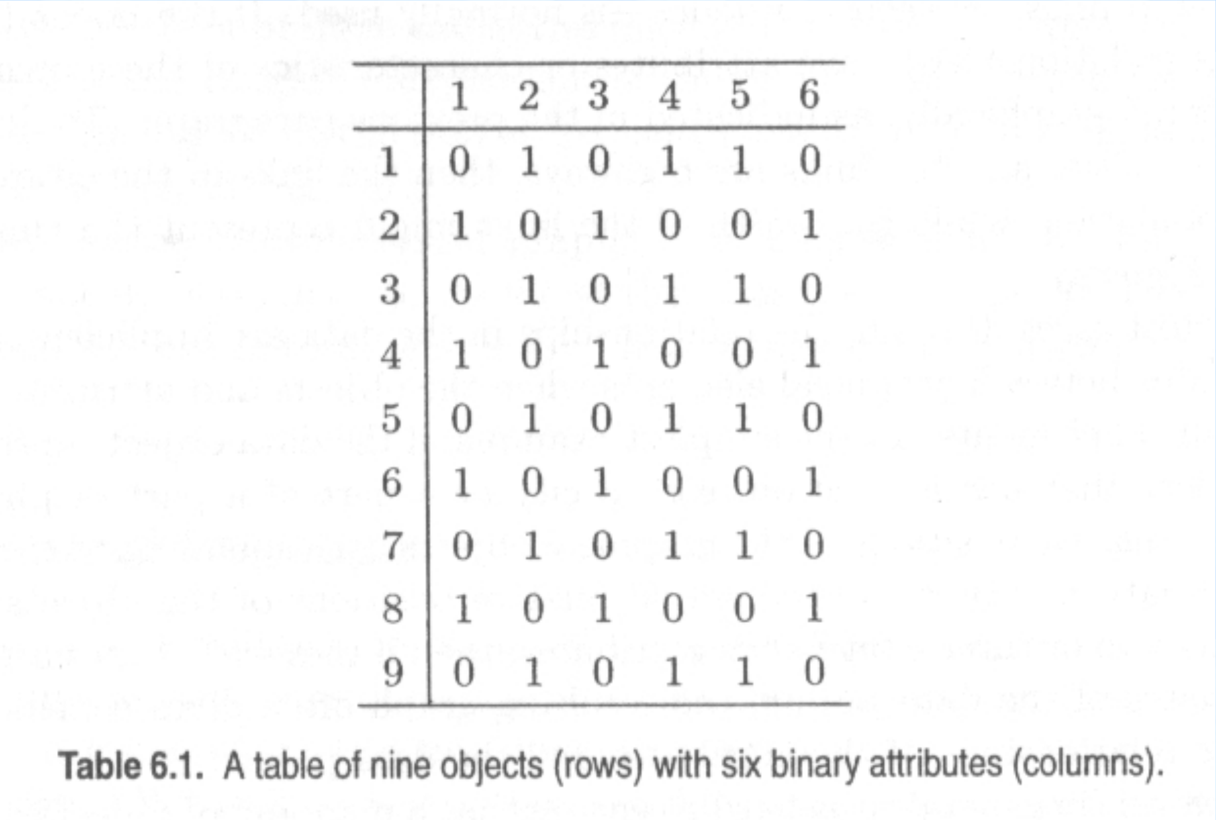
- Relationships are represented either explicitly or implicitly
- If we have graph data, then the standard graph representation – a set of nodes with links between the nodes, is normally used to show relationships
  - E.g. if the nodes are cities and the links are highways, then the linked nodes might represent connected cities, and width of the link might represent volume of traffic on that highway
- In most cases though the relationships in data are implicitly mapped to relationships between graphical objects e.g. relative positions, etc.

Representation …

- In general it is hard to ensure that a mapping of objects and attributes will also result in the relationships being mapped to easily observable relationships among graphical objects
- Given a set of data, there are many implicit relationships in the data; thus a key task of visualization is to choose a technique that makes the relationships of interest easily visible

Arrangement

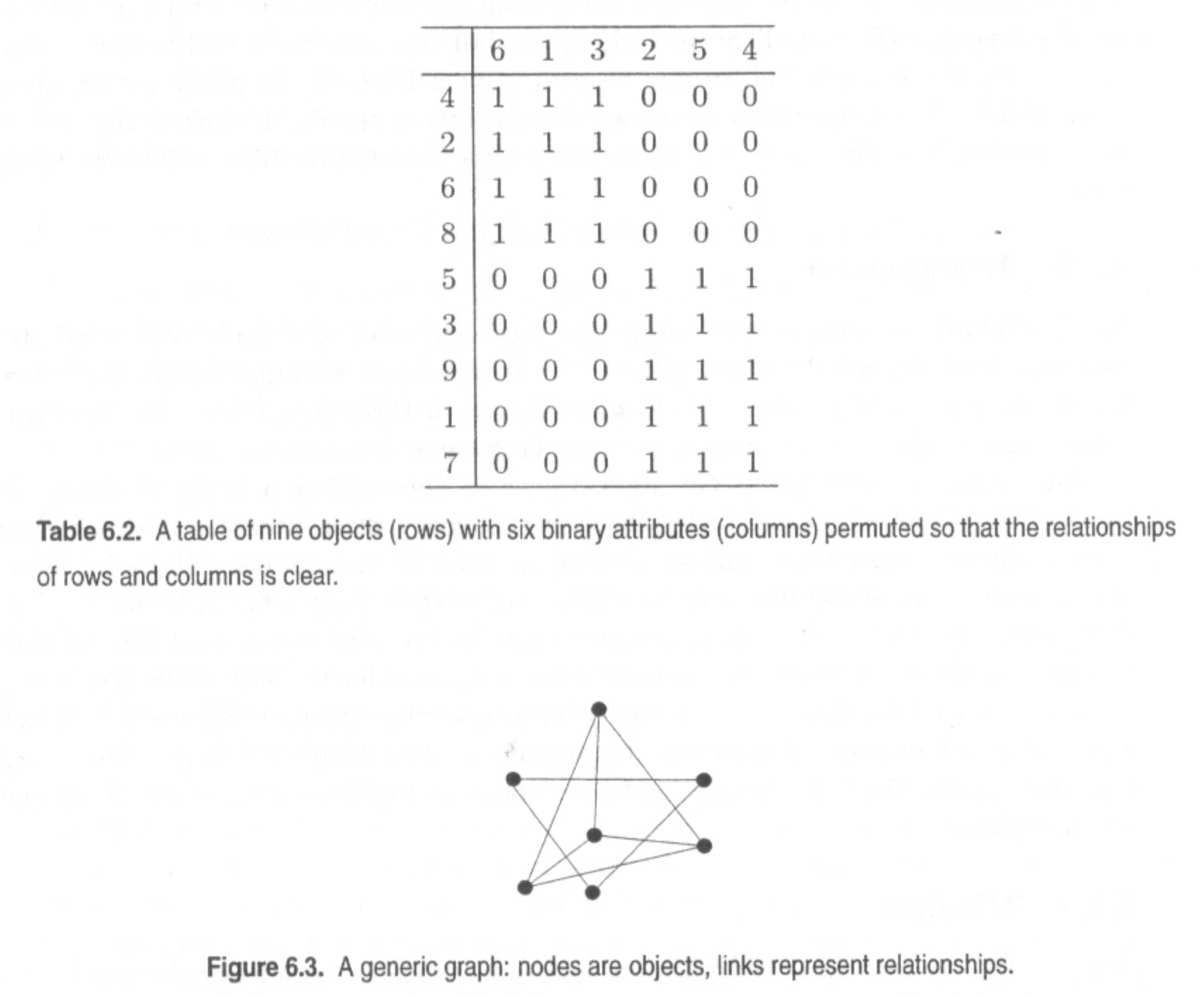
- Not only the representation, but also the arrangement of items within the visual display is crucial.
- E.g.



	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

Table 6.1. A table of nine objects (rows) with six binary attributes (columns).

Arrangement …



	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

Table 6.2. A table of nine objects (rows) with six binary attributes (columns) permuted so that the relationships of rows and columns is clear.

Figure 6.3. A generic graph: nodes are objects, links represent relationships.

Arrangement …

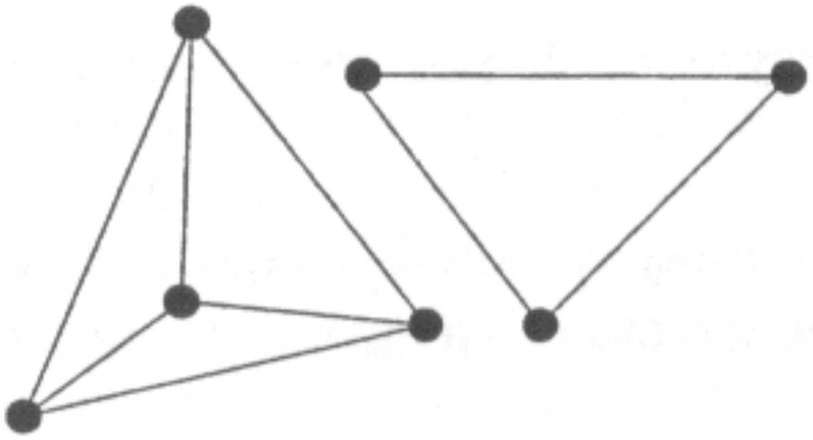


Figure 6.4. A generic graph: nodes are objects, links represent relationships.

Selection

- Elimination or de-emphasis of certain objects or attributes
- Projection – only a subset of attributes (usually two) is selected for display
- Sampling
- Zooming in on a certain region of data set

Do's and Don'ts

- ACCENT principles by D. A. Burn
  - **Apprehension:** Ability to correctly perceive relations among variables
  - **Clarity:** Ability to visually distinguish all the elements of a graph
  - **Consistency:** Ability to interpret a graph based on similarity to previous graphs
  - **Efficiency:** Ability to portray a possibly complex relation in as simple a way as possible
  - **Necessity:** The need for the graph, the graphical elements
  - **Truthfulness:** Ability to determine the true value represented by any graphical element using its magnitude relative to the implicit or explicit scale

Do's and Don'ts …

- Principles for graphical excellence by Edward R. Tufte
  - Graphical excellence is a well designed presentation of interesting data – a matter of substance, of statistics, and of design
  - Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency
  - Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest place
  - Graphical excellence is nearly always multivariate
  - Graphical excellence requires telling the truth about data

Visualization Techniques

- We talk about various one-dimensional visualization techniques using the following data set:
  - 50 iris flowers of three types each.
  - Each flower has five attributes
    - Sepal length (cm)
    - Sepal width (cm)
    - Petal length (cm)
    - Petal width (cm)
    - Class (Iris Setosa, Iris Versicolour, Iris Virginica)

Stem and Leaf Plots

- Stem and leaf plots can provide insight in to the distribution of one dimensional integer or continuous data
- In the simplest type of stem and leaf plot, we split the values in groups, where each group contains those values that are the same except for the last digit
- Another approach is to make buckets within each such group

Stem and Leaf Plots …

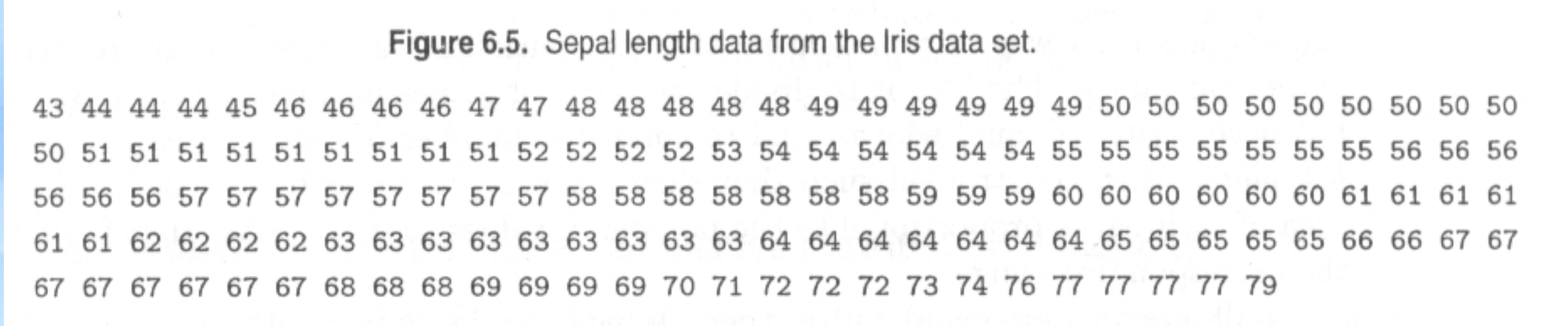


Figure 6.5. Sepal length data from the Iris data set.

```
43 44 44 44 45 46 46 46 46 47 47 48 48 48 48 48 49 49 49 49 49 49 50 50 50 50 50 50 50 50 50
50 51 51 51 51 51 51 51 51 51 52 52 52 52 53 54 54 54 54 54 54 55 55 55 55 55 55 55 56 56 56
56 56 56 57 57 57 57 57 57 57 57 58 58 58 58 58 58 58 59 59 59 60 60 60 60 60 60 61 61 61 61
61 61 62 62 62 62 63 63 63 63 63 63 63 63 63 64 64 64 64 64 64 64 65 65 65 65 65 66 66 67 67
67 67 67 67 67 67 68 68 68 69 69 69 69 70 71 72 72 72 73 74 76 77 77 77 77 79
```

Stem and Leaf Plots …

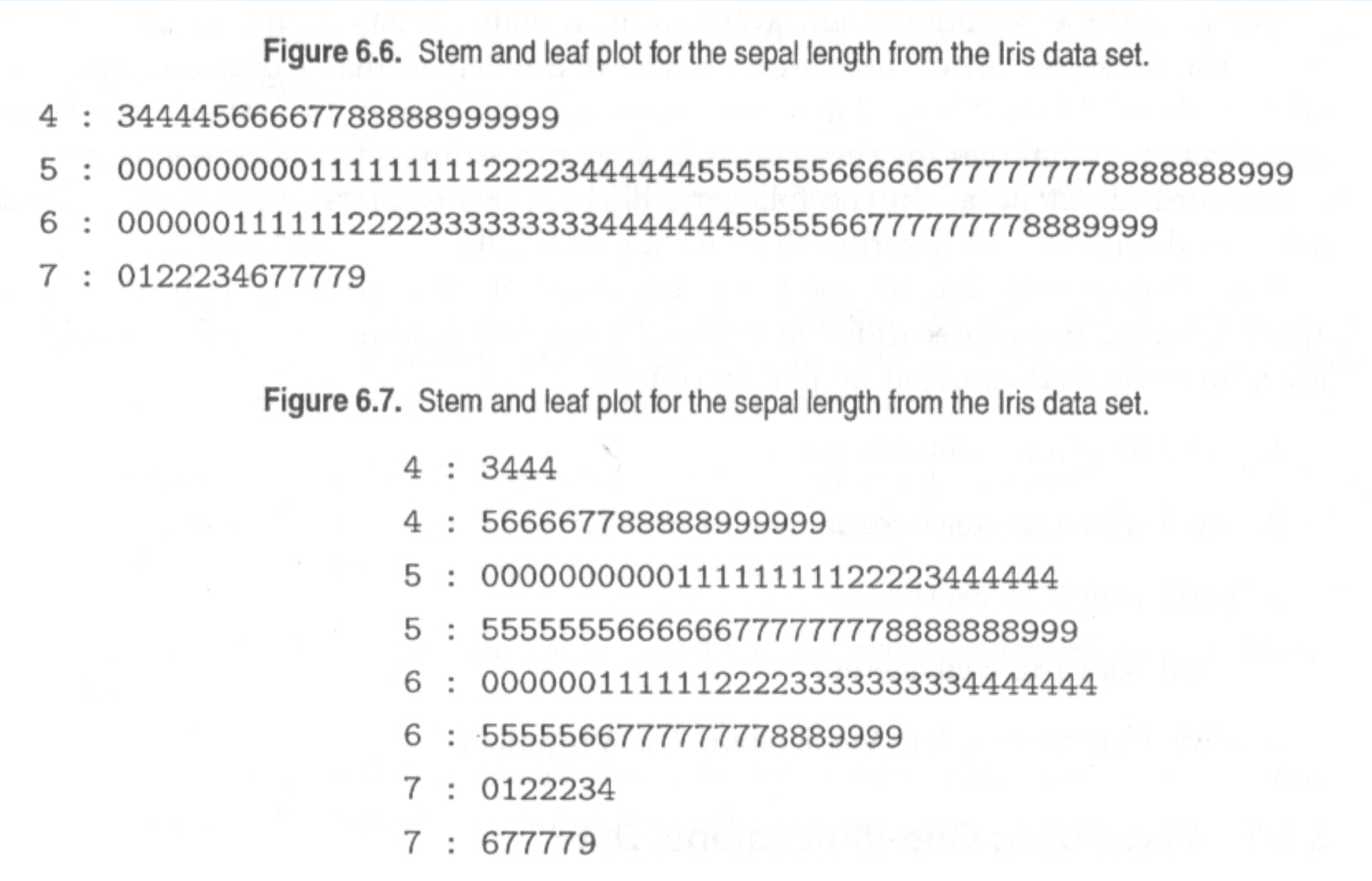


Figure 6.6. Stem and leaf plot for the sepal length from the Iris data set.

```
4 : 3444566667788888999999
5 : 0000000000111111111222234444445555555666666777777778888888999
6 : 000000111111222233333333344444445555566777777778889999
7 : 0122234677779
```

Figure 6.7. Stem and leaf plot for the sepal length from the Iris data set.

```
4 : 3444
4 : 566667788888999999
5 : 000000000011111111122223444444
5 : 5555555666666777777778888888999
6 : 00000011111122223333333334444444
6 : 5555566777777778889999
7 : 0122234
7 : 677779
```

Dot Plots

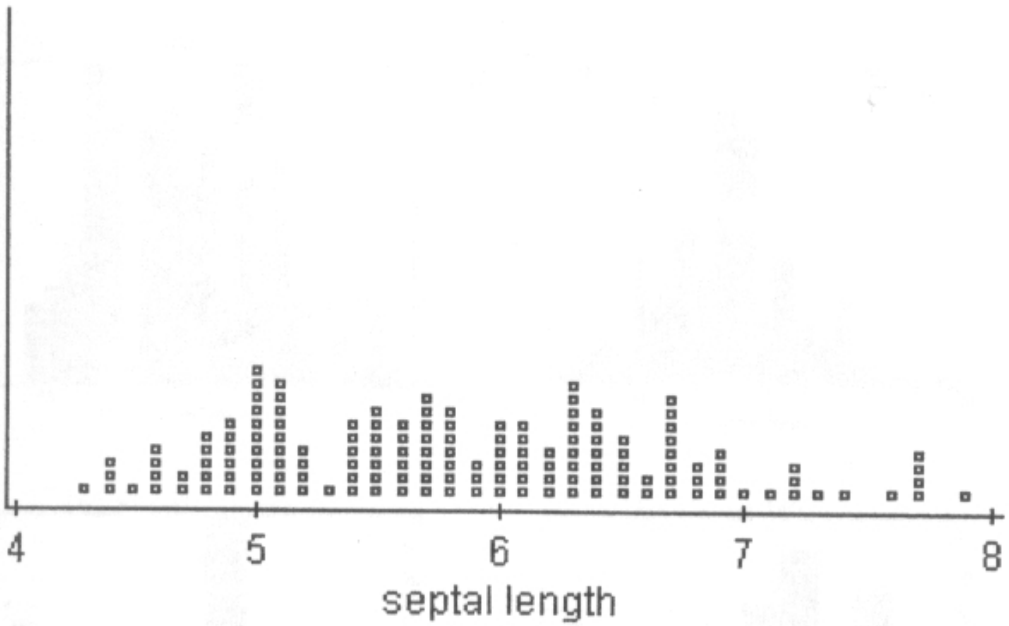


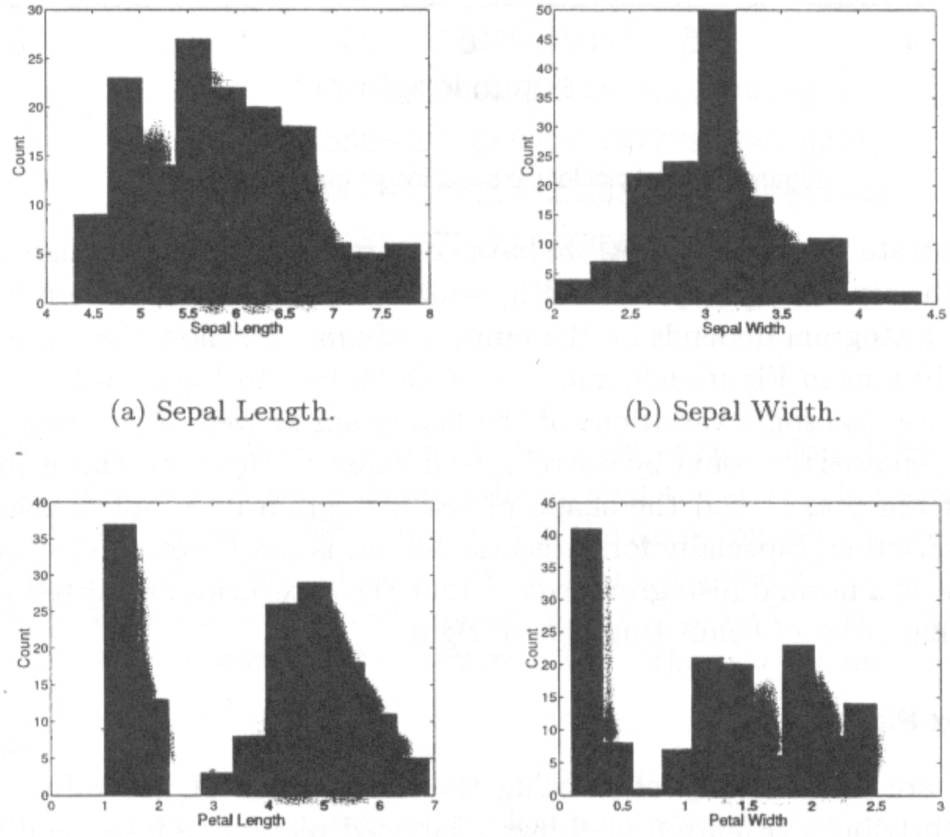
Figure 6.8. Dot plot for the sepal length from the Iris data set

- A dot plot is a variation on a stem and leaf plot, where we do not show the digits in each bucket, but rather, only show a dot for each data object that falls in each bucket

Histograms

- Way of representing the distribution of values for numerical attributes
- Divide the range of values in to bins, and count the number of values in each bin
- Construct a bar plot
  - Each bin is represented by a bar
  - Area of bar is proportional to the count
- If attribute is categorical, then often one bin is allotted to each value

Histograms …



(a) Sepal Length.

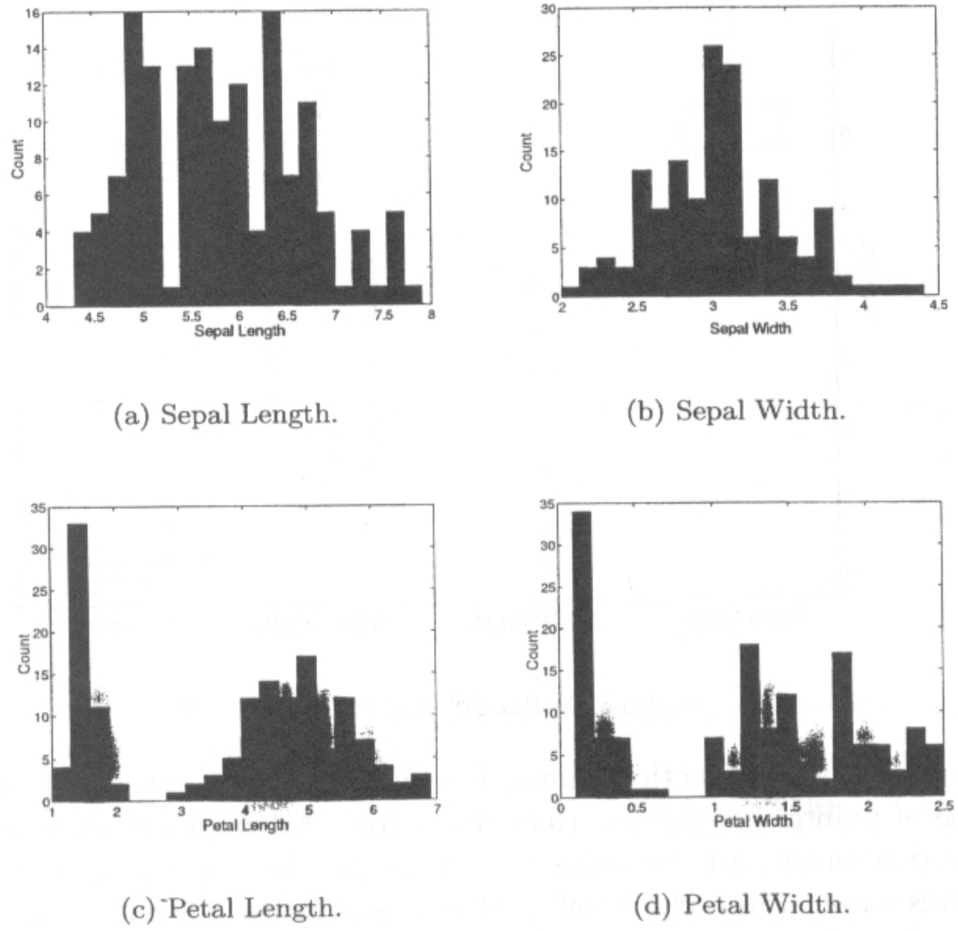
(b) Sepal Width.

(c) Petal Length.

(d) Petal Width.

Figure 6.9. Histograms of Four Iris Attributes - 10 bins.

Histograms …



(a) Sepal Length.

(b) Sepal Width.

(c) Petal Length.

(d) Petal Width.

Figure 6.10. Histograms of Four Iris Attributes - 20 bins.

Histograms …

- A variation of histogram is to replace the count by relative frequency
- Another common variation, especially for categorical data, is the Pareto Histogram, which is the same as a normal histogram except that the categories are sorted horizontally in decreasing order of count from left to right

Box plots

- Another way of showing the distribution of the values of a single numerical attribute
- Lower and upper ends of the box represent 25th and 75th percentile. The line inside the box represents 50th percentile
- Top and bottom lines indicate 10th and 90th percentile
- Outliers are shown by '+' marks

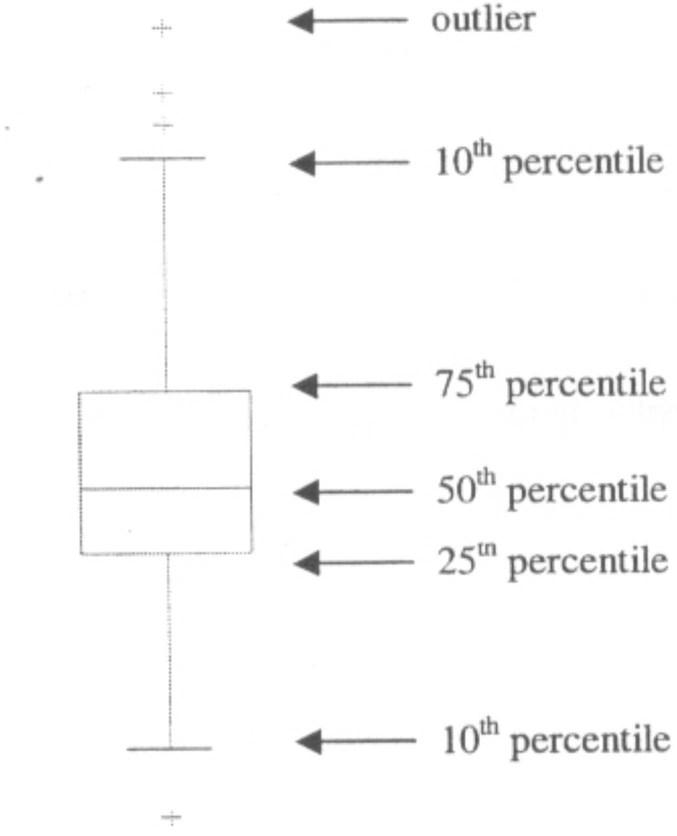


Figure 6.11. Description of Box Plot.

Box Plots ...

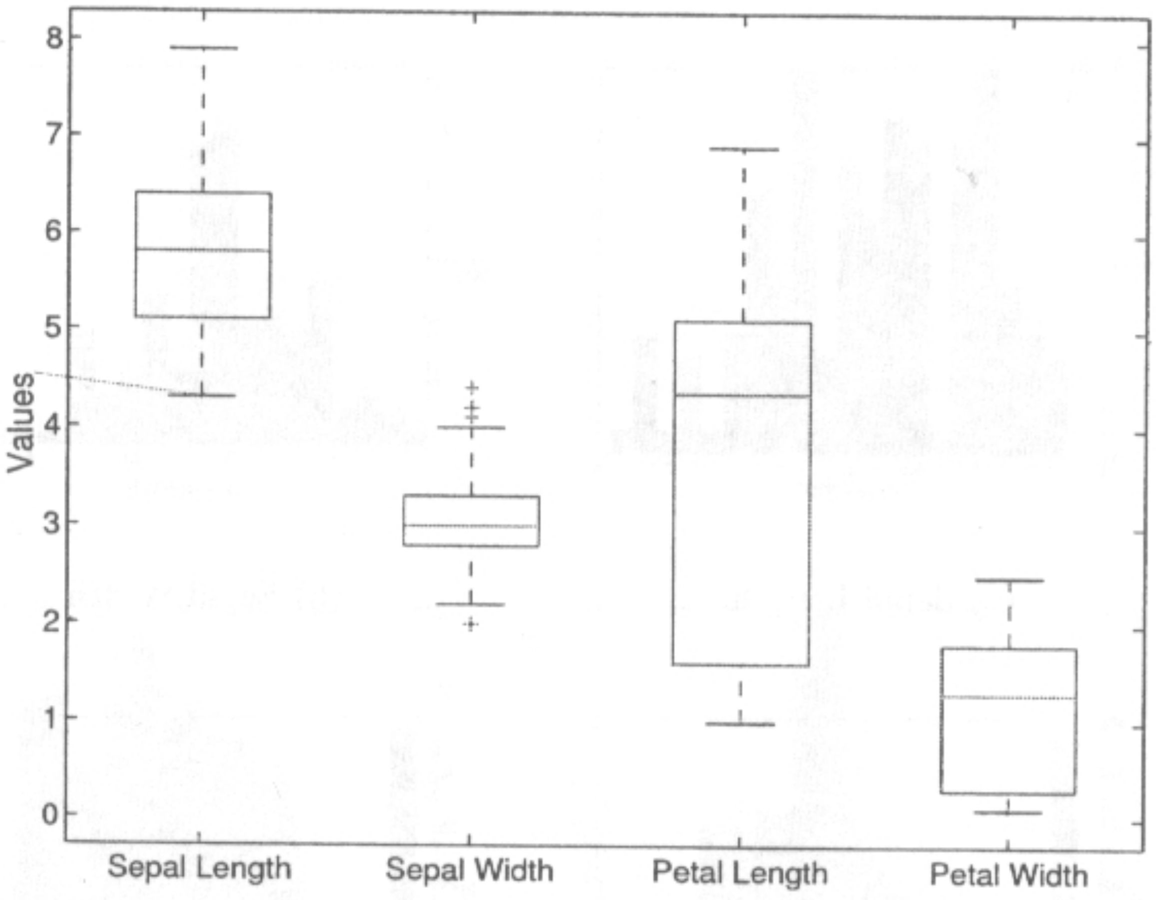
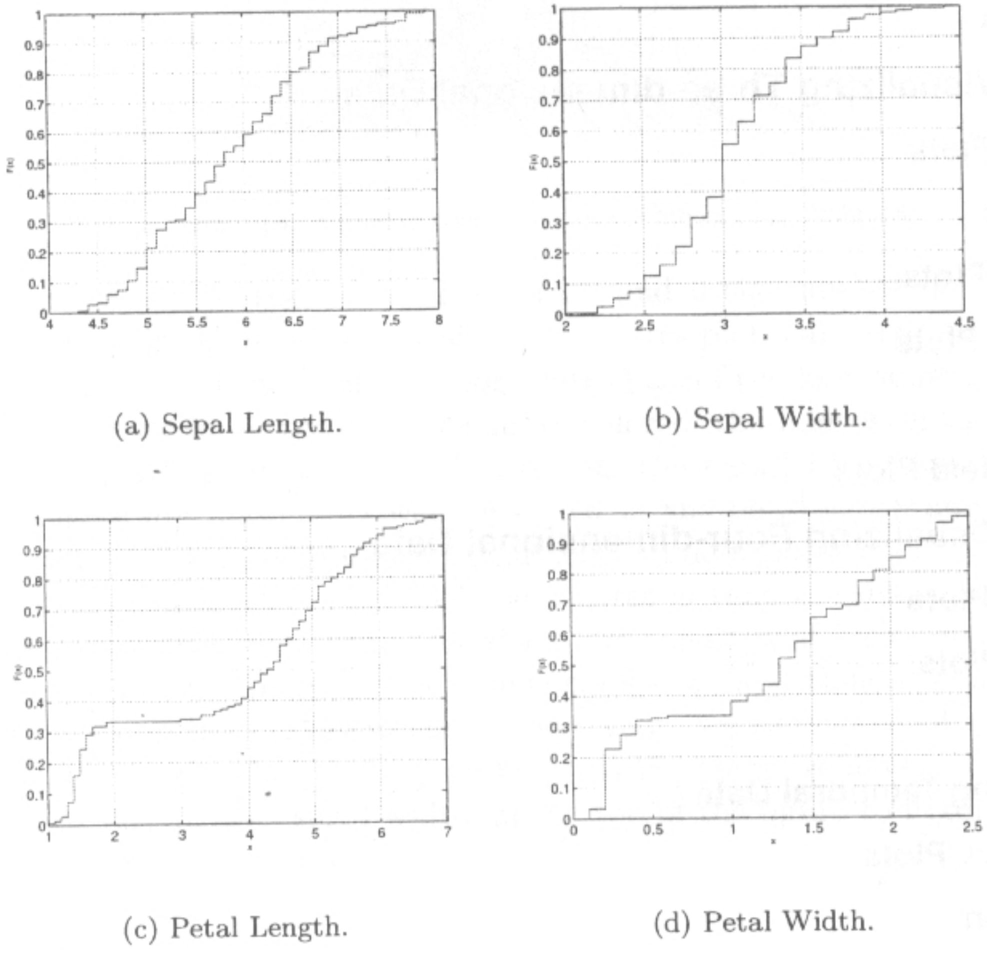


Figure 6.12. Box plot for Iris attributes.

Percentile Plots and Empirical Cumulative Distribution Functions

- A cumulative distribution function shows, for each value of a statistical distribution , the fraction of points that are less than that value
- An empirical cumulative distribution function shows, for the value of each point, the fraction of points that are less than this value. Since the number of points is finite, the empirical cumulative distribution function is a step function

Empirical Cumulative Distribution Functions …



(a) Sepal Length.

(b) Sepal Width.

(c) Petal Length.

(d) Petal Width.

Figure 6.13. Empirical CDF's of Four Iris Attributes.

Visualizing Higher-Dimensional Data

- Parallel Coordinates
  - An approach for displaying data that have many attributes
  - We have one coordinate axis for each attribute
  - Instead of placing coordinate axes perpendicular to each other, we place them parallel to each other
  - Instead of representing an object as a point, we represent it as a line i.e. for each value of an attribute of an object, we put a point on the corresponding coordinate axis value and then connect all of these points

Parallel Coordinates Plot

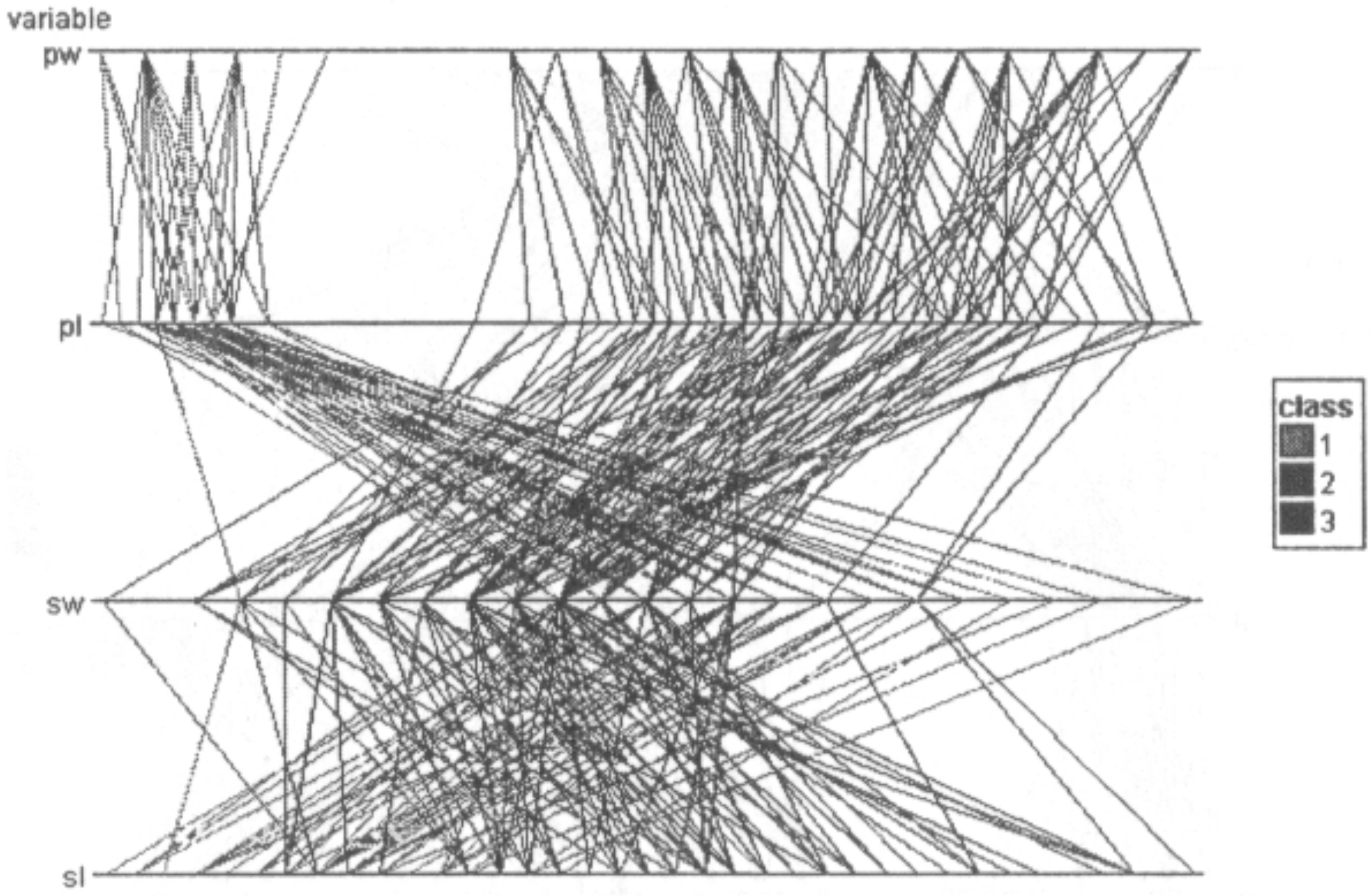


Figure 6.14. A parallel coordinates plot of the four iris attributes.

Parallel Coordinates Plot …

- One of the drawbacks of parallel coordinates plot is that the detection of patterns depends upon the order of the axes
- If the lines are crossing over one another a lot, the picture can become confusing
- It is desirable to order the axes in such a way that we have minimum crossover

Parallel Coordinates Plot …

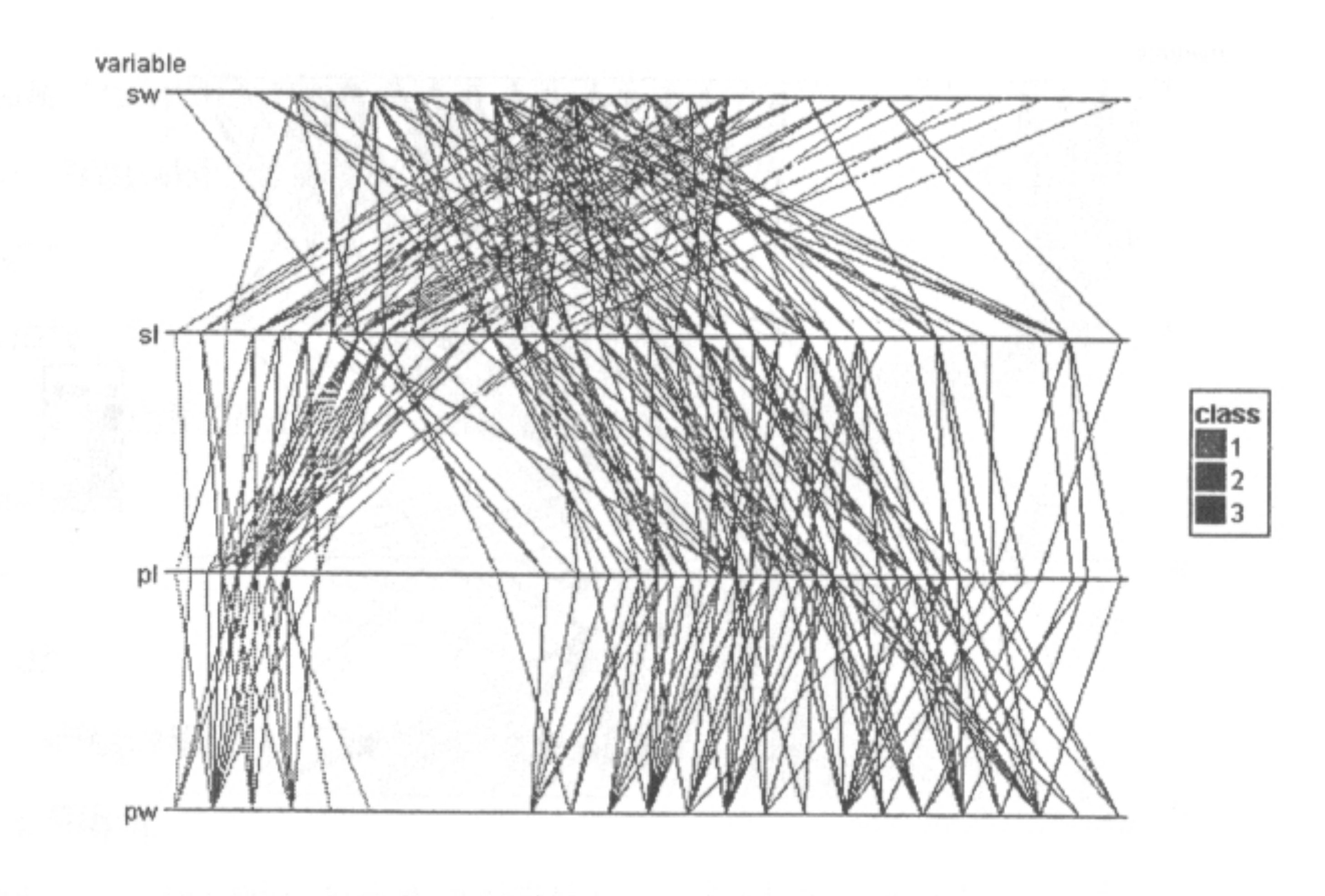


Figure 6.15. A parallel coordinates plot of the four iris attributes.