Data Mining

Exploring Data

Slides by Tan, Steinbach, Kumar adapted by Pimprapai Thainiam

What is Data Exploration?

Data exploration is a preliminary exploration of the data to better understand its characteristics by using summary statistics and data visualization.

- **Key motivations** of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns from visualization because people can recognize patterns that is not captured by data analysis tools

- **Two techniques** used in data exploration:
  1. Summary statistics
  2. Visualization

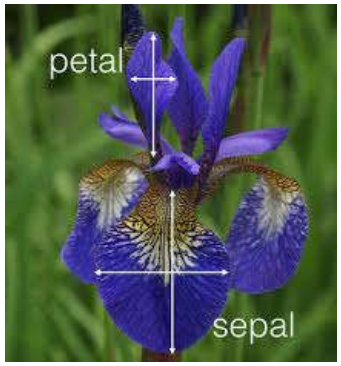
Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository or Package Data in R
  - The dataset is contributed by the statistician named Douglas Fisher
  - Three flower types (species):

Setosa

Virginica

Versicolour



  - It consists of 150 flowers with 4 continuous attributes and 1 categorical attribute
    1) Sepal width in cm. 2) Sepal length in cm.
    3) Petal width in cm. 4) Petal length in cm.
    5) Class → {Setosa, Vesicolour, Virginica}

Topics

- **Summary Statistics**
- Visualization Techniques
  - Visualizing Small Numbers of Attributes
  - Visualizing Higher-Dimensional Data

Summary Statistics

Summary statistics are quantities that capture various characteristics of a potentially large set of values with a single number or a small set of numbers used to describe the data observations in various aspects, such as

1. **Measure of frequency**, such as frequency and mode
2. **Measure of location** (aka central tendency) such as the arithmetic mean and median
3. **Measure of spread**, or statistical dispersion, such as the standard deviation
4. **Measure of statistical dependence between attributes** such as a correlation coefficient; this measure is used when more than one attributes is measured

Measures of Frequency: Frequency and Mode

- **Frequency** of a categorical 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.
  - Given a categorical attribute x, which takes on values $\{v_1, \ldots, v_i, \ldots, v_k\}$ and a set of m objects, the frequency of a value v_i is defined as

$$frequency\,(v_i) = \frac{number\ of\ objects\ with\ attribute\ value\ v_i}{m}$$

- **Mode** of a categorical attribute is the value that has the highest frequency.
- The notions of frequency and mode are typically used with categorical data.

att A
1
1
1
2
2
2
3
3
3
3

Mode = 3

$f_1 = \frac{3}{10}$, $f_2 = \frac{3}{10}$, $f_3 = \frac{4}{10}$

Percentiles and Quantiles

- **Percentiles:** For ordered data, it is more useful to consider the percentiles of a set of values.

- 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.
  - The 0^{th} percentile is equal to minimum
  - The 25^{th} percentile is known as the lower quartile
  - The 50^{th} percentile is known as the median.
  - The 75^{th} percentile is known as the upper quartile.
  - The 100^{th} percentile is equal to maximum

- **Quantiles** are the same as percentiles, but are indexed by fractions rather than by percentages.

Percentiles and Quantiles

Steps to find percentile and quantile

Given a set of values = {3.7, 2.7, 3.3, 1.3, 2.2, 3.1} , find 25^{th} and 50^{th} percentiles.

1. Sort the values into order:

1.3 2.2 2.7 3.1 3.3 3.7

2. Associate the ordered values with sample percentiles (fractions) equally spaced from 0 to 100 (0 to 1).

Percentile →

Sample percentile	0	20	40	60	80	100
Quantiles	1.3	2.2	2.7	3.1	3.3	3.7

Quantile →

Sample fraction	0	0.2	0.4	0.6	0.8	1.0
Quantiles	1.3	2.2	2.7	3.1	3.3	3.7

3. Calculate quantiles by linear interpolation between the two values.

→ percentiles or

Percentiles and Quantiles

Sample percentile	0	20	40	60	80	100
Quantiles	1.3	2.2	2.7	3.1	3.3	3.7

- The lower quartile (25th percentile, quantile 0.25) corresponds to a sample percentile of 25. The lower quartile must be

$$\frac{x - 2.2}{2.7 - 2.2} = \frac{25 - 20}{40 - 20}$$

$$x = 2.325$$

- The median (50th percentile, quantile 0.5) corresponds to a sample percentile of 50. The median must be

$$\frac{x - 2.7}{3.1 - 2.7} = \frac{50 - 40}{60 - 40}$$

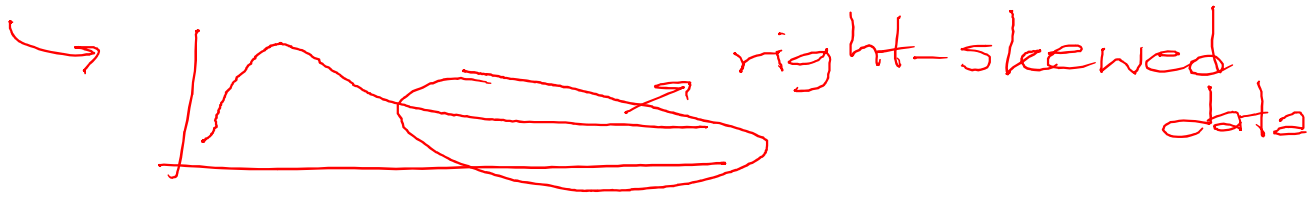
$$x = 2.9$$

Measures of Location: Mean and Median

- **Mean** is the most common measure of the location of a set of values.
- Consider a set of m objects and an attribute x; let $\{x_1, \dots, x_m\}$ be the attribute values of x for these m objects.

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

- Mean is very sensitive to outliers.
- Mean is interpreted as the middle of a set of values, this is only correct if the values are distributed in a symmetric manner.
- If the distribution of values is skewed, then the median is a better indicator of the middle.



Measures of Location: Mean and Median

- **Median** can be used as a measure of the location when there are some outliers in the data set.

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

location of median

where m is sample size or number of objects

- Example: $x = \{11,12,13,14,15,16,17,18,19,200\}$ → outlier

$m = 10$

$10 = 2r$

$r = 5$

$$mean = \frac{335}{10} = 33.5$$

$$median = \frac{15+16}{2} = 15.5$$

$median = \frac{1}{2}(x_5 + x_6)$

Measures of Location: Mean and Median

- **Trimmed mean** can also be used as a measure of the location when there are some outliers in the data.
- Given a percentage p between 0 and 100; the top and bottom $\frac{p}{2}$% of the data is thrown out, and the mean is then calculated in the normal way.
- The standard mean is a trimmed mean with p = 0%
- For example, p = 20 $\rightarrow \frac{p}{2}$ = 10; exclude 10% in the top and 10% in the bottom of the data point; so, exclude 11 and 200 from mean calculation.

$$x = \{11,12,13,14,15,16,17,18,19,200\}$$

$$trimmed\ mean = \frac{12 + 13 + 14 + 15 + 16 + 17 + 18 + 19}{8} = 15.5$$

Measures of Spread: Range and Variances

- **Range** is the difference between the maximum and minimum

- Given an attribute x, with a set of m ordered values $\{x_1, \dots, x_m\}$

$$range(x) = \max(x) - \min(x) = x_m - x_1$$

- The range identifies the maximum spread, it can be misleading if there are some outliers in the data.

Measures of Spread: Range and Variances

- **Variance (s^2)** measures how far a set of values are spread out from their average value. It is the most common measure of the spread of a set of points.

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

m = Number of objects

$\bar{x}$ → mean

- However, variance is also sensitive to outliers, so that more robust measures are often used.

Average Absolute Deviation:

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

Median Absolute Deviation:

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

Interquartile Range:

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

Multivariate Summary Statistics

- **Covariance** between attributes i and j is a measure of the degree to which two attributes vary together and depends on the magnitudes of the variables.

$$s_{ij} = covariance(x_i, x_j) = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ki} - \bar{x_i})\left(x_{kj} - \bar{x_j}\right)$$

where s_{ij} is the covariance of the i^{th} and j^{th} attributes of the data

x_{ki} and x_{kj} are the values of the i^{th} and j^{th} attributes for the k^{th} object

$\bar{x_i}$ and $\bar{x_j}$ are the means of the i^{th} and j^{th} attributes

m is number of object

- s_{ij} → 0 indicates that two attributes do not have a (linear) relationship.

  → 1 indicates that two attributes have a (linear) relationship.
- Note that $covariance(x_i, x_i) = variance(x_i)$

Multivariate Summary Statistics

- **Correlation** between attributes i and j gives an indication of how strongly two attributes are (linearly) related.

- For example, height and weight are related; taller people tend to be heavier than shorter people.

considered as normalized covariance

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

covariance

where s_i, s_j are the variance of attribute i and j, respectively

- Note that correlation is standardized covariance, thus correlation value is range between -1 and 1.

Topics

- Summary Statistics
- **Visualization Techniques**
  - Visualizing Small Numbers of Attributes
  - Visualizing Higher-Dimensional Data

High Numbers of Attributes

Visualization

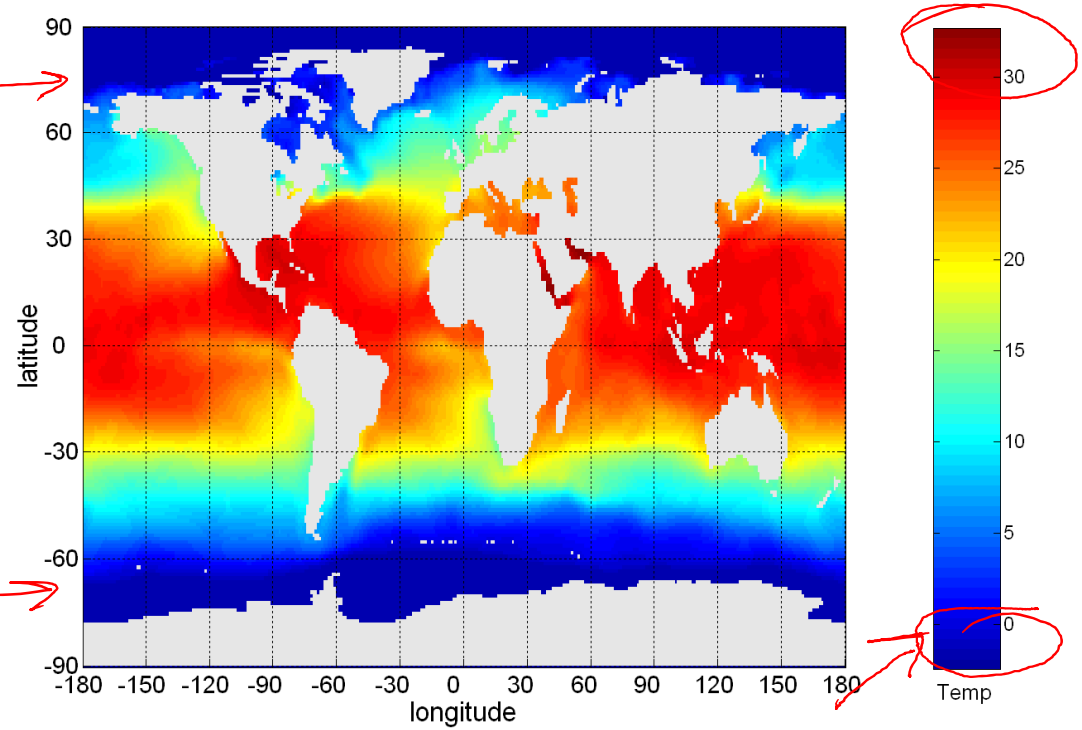
- **Visualization** is the conversion of data into a graphic or tabular (table) format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- Visualization of data is 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

Visualization

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) in degree Celsius for July 1982
- It summarizes the information from approximately 250,000 of data points in a single figure.



Visualization

The three key concept in visualization:

1. Representation
2. Arrangement
3. Selection

Representation

- **Representation** is the step of mapping information to a visual format.

- A set of information consists of data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.

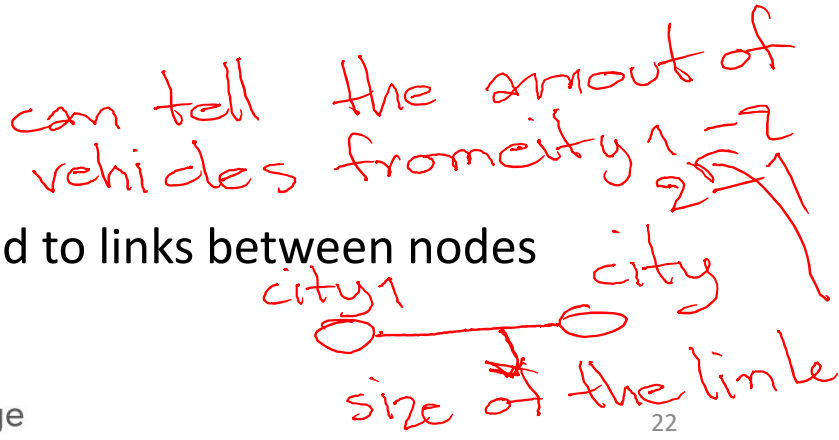
Visualization

- Example for data objects
    - Objects of a single attribute are often represented as
        - entry in a table
        - an area on screen
    - Objects of multiple attributes are often represented as
        - a row or column of a table
        - a line on a graph
        - a point in two- or three- dimensional space

Visualization

- Example for attributes
  - Ordinal and continuous attributes can be mapped to continuous, ordered graphical features, such as
    - location along x, y, and z axes
    - color
    - size (diameter, width, height, etc.)
  - Categorical attributes, each category can be mapped to
    - a distinct position, color, shape, orientation, embellishment
    - a column in a table

- Example for relationships
  - Relationships between objects can be mapped to links between nodes (objects).



Visualization

Arrangement

- **Arrangement** is the placement of visual elements within a display so that the patterns and trends can be observed easily.

- The proper choice of visual representation of objects and attributes 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

Visualization

Selection

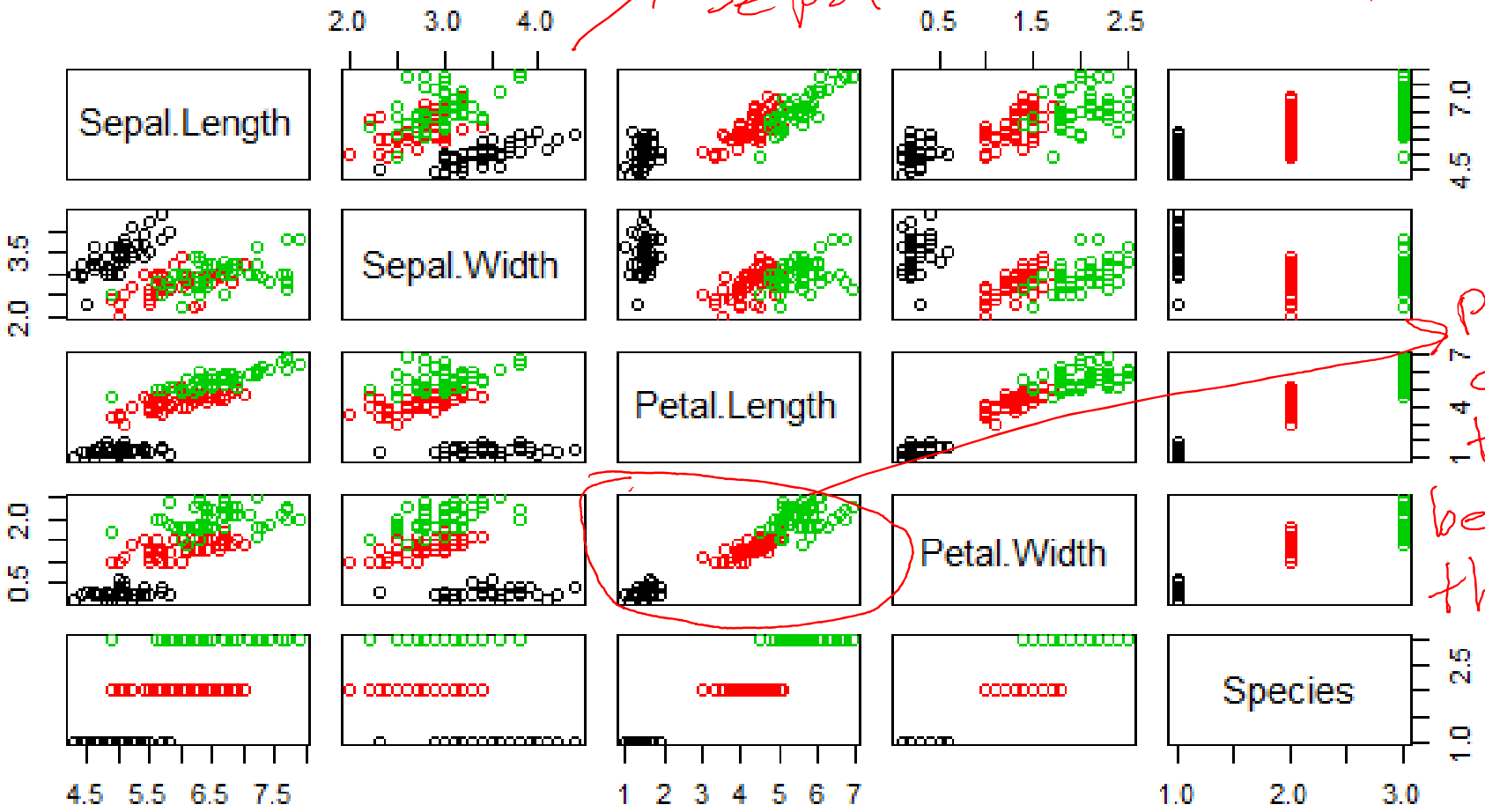
- The most common approach to handling many attributes (data objects) is to choose a subset of attributes (data objects) to display.

- Many attributes:
  - Choose a subset of attributes (2 or 3 attributes) to display
  - If the number of attributes (dimensionality) is not too high, a matrix of bivariate (two-attribute) plots can be constructed.

- Many data objects:
  - Zooming in on a particular region of the data
  - Taking a sample of the data points using a sample technique

Visualization

Bivariate (two-attribute) plots



sepal length (y axis)
Sepal width (x axis)

possitive correlation between these 2 attributes

Topics

- Summary Statistics
- Visualization Techniques
  - **Visualizing Small Numbers of Attributes**
  - Visualizing Higher-Dimensional Data

Visualizing Small Numbers of Attributes

Stem and Leaf Plots

- It can be used to provide insight into the distribution of one-dimensional integer or continuous data.
- **Steps:**
    - Sort the data objects from smallest to largest
    - Split the value into groups, where each group (stem) contains those values that are the same except for the last digits (leave)
        - The stems will be the high-order digits (first n-1 digits)
        - The leaves will be the low-order digits (last digits)
    - Plot the stems vertically and leaves horizontally, this can provide a visual representation of the distribution of the data

- Note: Stem and leaf plots are a type of histogram

Visualizing Small Numbers of Attributes

- **Example:** Sepal length of Iris data set

```
4.3 4.4 4.4 4.4 4.5 4.6 4.6 4.6 4.6 4.7 4.7 4.8 4.8 4.8 4.8 4.8 4.9 4.9 4.9 4.9 4.9 4.9 5.0 5.0 5.0 5.0 5.0
5.0 5.0 5.0 5.0 5.0 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.1 5.2 5.2 5.2 5.2 5.3 5.4 5.4 5.4 5.4 5.4 5.4 5.5 5.5
5.5 5.5 5.5 5.5 5.5 5.6 5.6 5.6 5.6 5.6 5.6 5.7 5.7 5.7 5.7 5.7 5.7 5.7 5.7 5.8 5.8 5.8 5.8 5.8 5.8 5.8 5.9
5.9 5.9 6.0 6.0 6.0 6.0 6.0 6.0 6.1 6.1 6.1 6.1 6.1 6.1 6.2 6.2 6.2 6.2 6.3 6.3 6.3 6.3 6.3 6.3 6.3 6.3 6.3
6.4 6.4 6.4 6.4 6.4 6.4 6.4 6.5 6.5 6.5 6.5 6.5 6.6 6.6 6.7 6.7 6.7 6.7 6.7 6.7 6.7 6.7 6.8 6.8 6.8 6.9 6.9
6.9 6.9 7.0 7.1 7.2 7.2 7.2 7.3 7.4 7.6 7.7 7.7 7.7 7.7 7.9
```

1 bin for each stem →

```
4 | 344456666778888899999
5 | 000000000011111111122223444444555555566666777777788888899 9
6 | 000000111111222233333333344444445555566777777778889999
7 | 0122234677779
```

2 bins for each stem →

```
4 | 3444
4 | 566667788888999999
5 | 00000000001111111122223444444
5 | 55555556666667777777888888999
6 | 00000011111122223333333334444444
6 | 555556677777778889999
7 | 0122234
7 | 677779
```

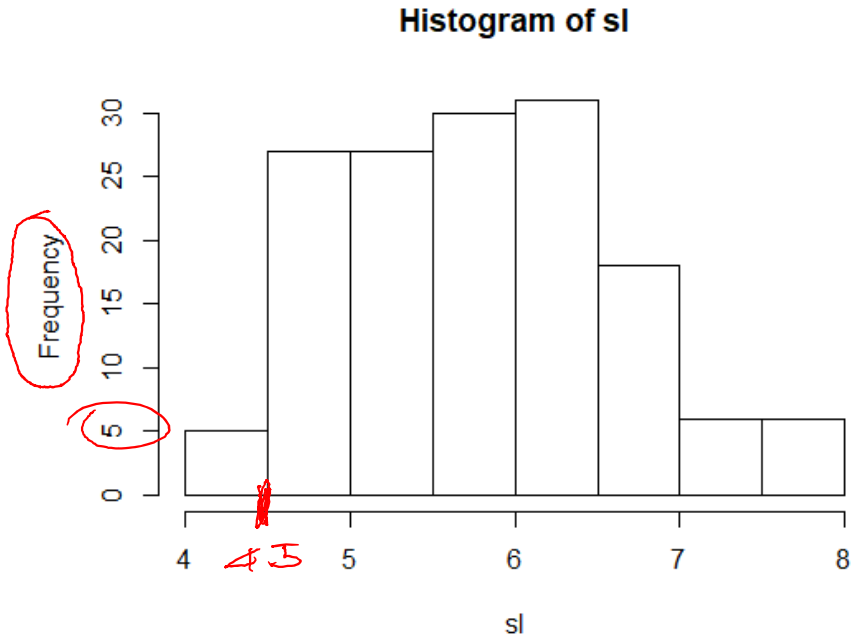
Visualizing Small Numbers of Attributes

Histogram

- Histogram is a plot that displays the distribution of values for attributes by dividing the possible values into bins and showing the number of objects that fall into each bin.
  - For categorical attributes, each value is a bin
  - For continuous attributes, the range of values is divided into bins (not necessary to be equal width)
- **Steps:**
  - Divide the values into bins if an attribute is continuous attribute; or map the values into bins if an attribute is categorical attribute
  - Count number of objects that fall into each bin
  - Create bar plot where each bin is represented by one bar and the area of each bar is proportional to the number of objects that falls into the corresponding range. For equal width bins, the height of each bar indicates the number of objects (count) that fall into each bin.

Visualizing Small Numbers of Attributes

- **Example:** Sepal length of Iris data set



4.3 – 7.9

4.0 – 8.0

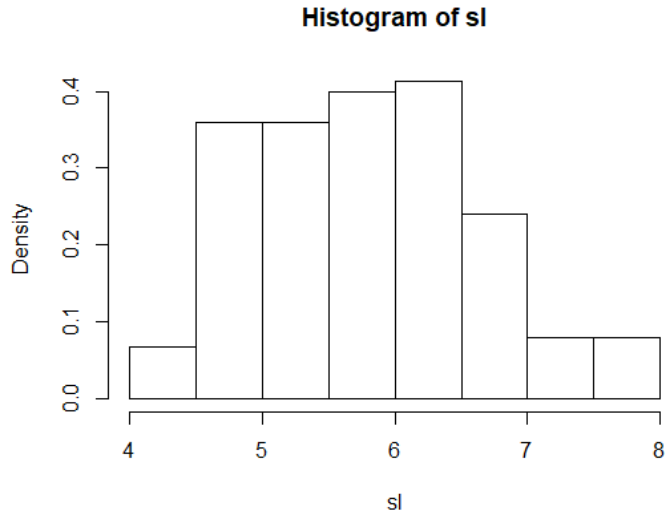
8 bins

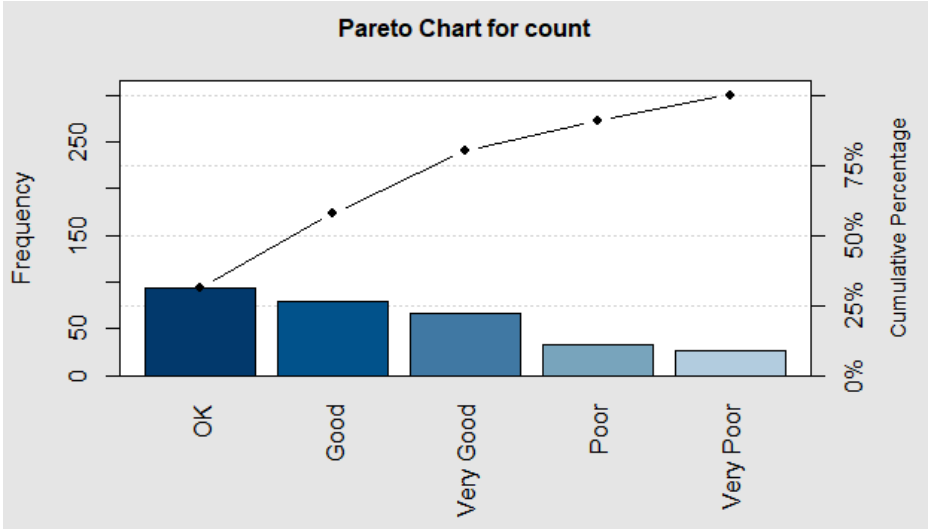
0.5 → width of each bin

Note: Frequency (y axis) = Count

Visualizing Small Numbers of Attributes

- **Histogram variations:**
  1. **Relative (frequency) Histograms:** It replaces the count by the relative frequency.
  2. **Pareto Histograms:** This type of histogram is used for unordered categorical data where the categories are sorted by count so that the count is decreasing from left to right.

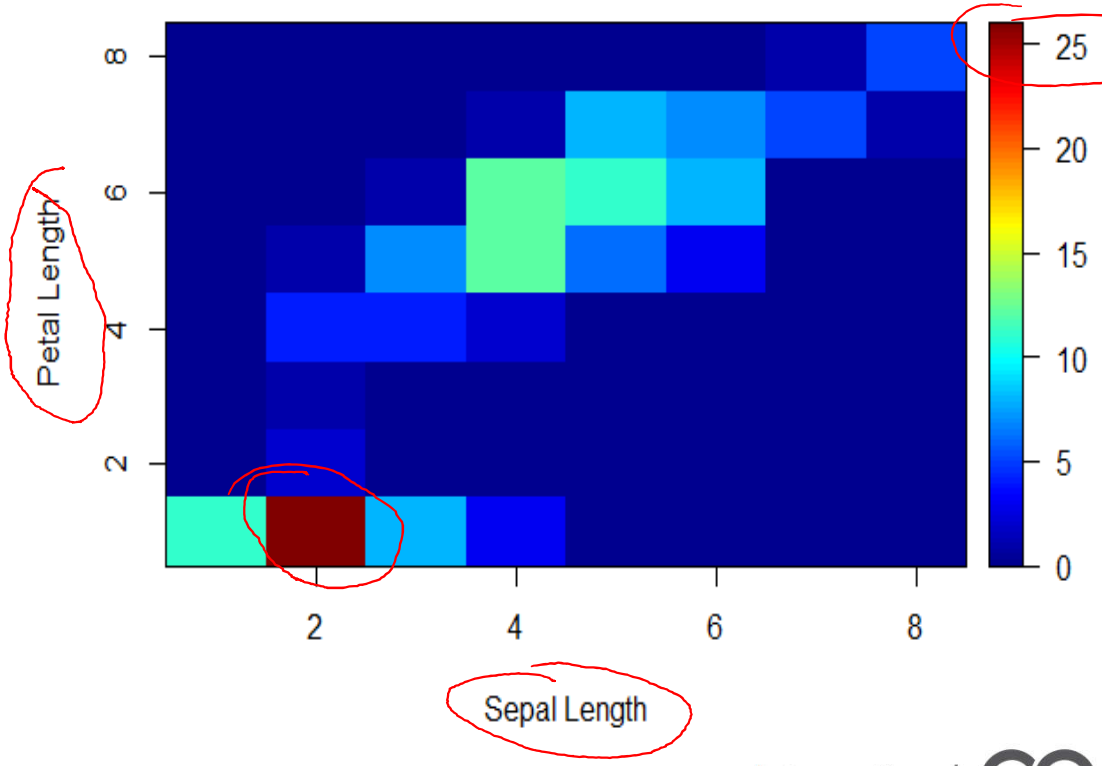


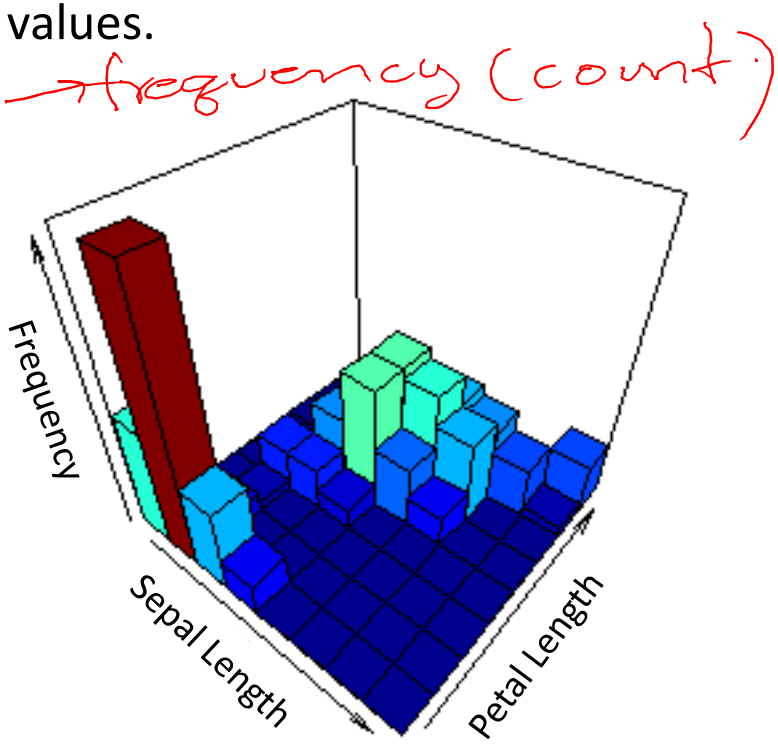


Note: Density = relative frequency

Visualizing Small Numbers of Attributes

3. Two-dimensional Histograms: Show the joint distribution of the values of two attributes where each attribute is divided into intervals and the two sets of intervals define two-dimensional rectangles of values.

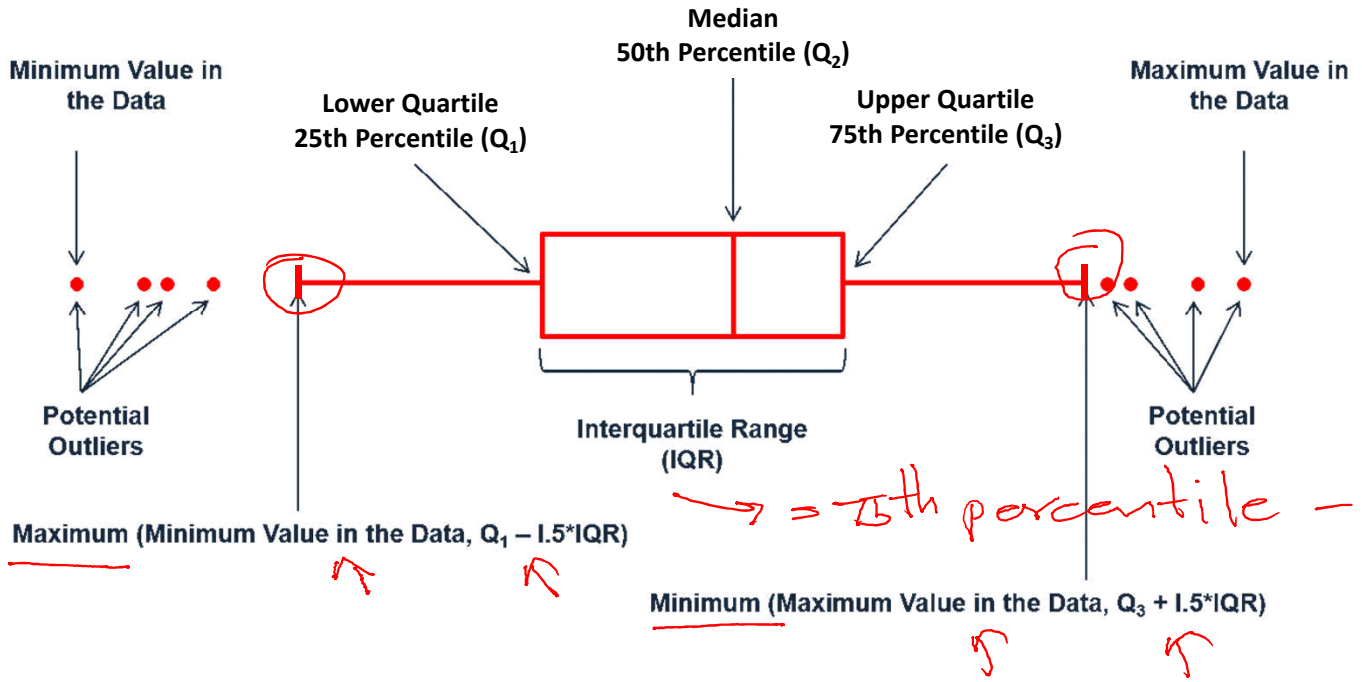




Visualizing Small Numbers of Attributes

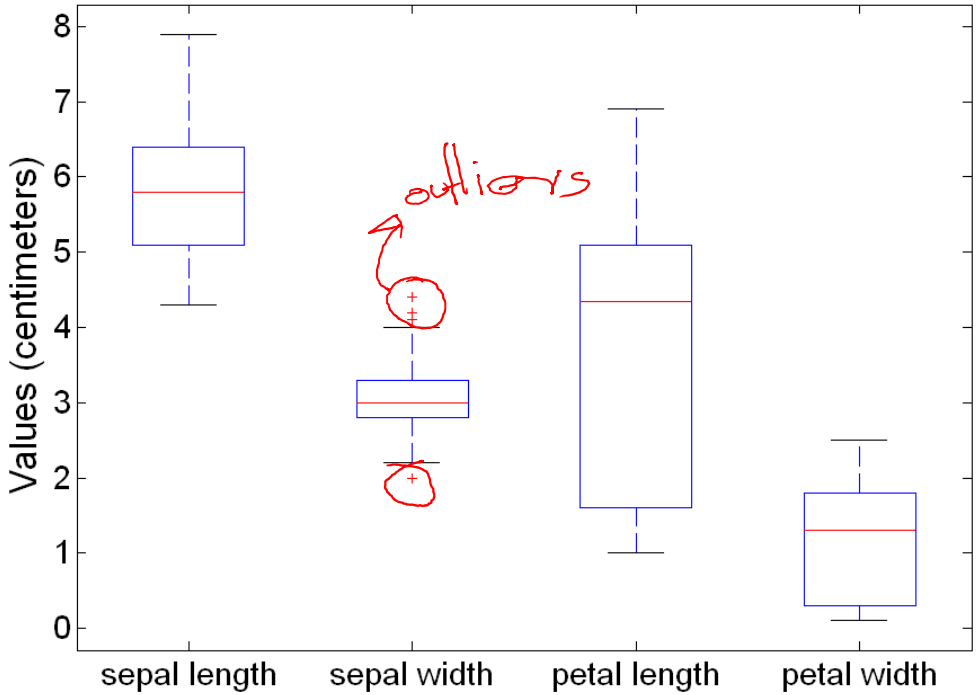
Box Plots

- Box Plots are another method for showing the distribution of the values of a single numerical attribute.



Visualizing Small Numbers of Attributes

- Box plots can be used to compare how attributes vary



Visualizing Small Numbers of Attributes

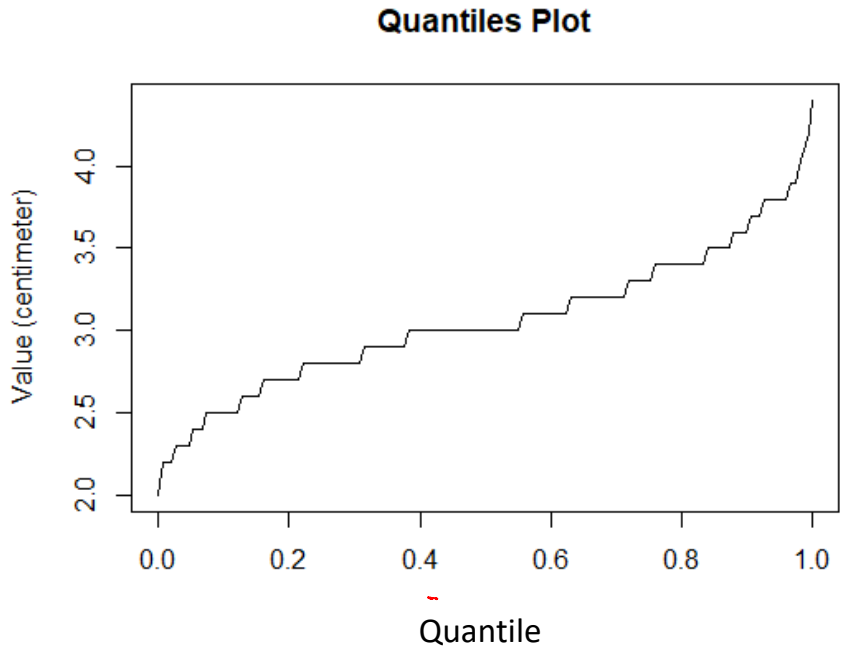
Pie Chart

- Pie chart is similar to histogram, but is typically used with categorical attributes that have relatively small number of values.
- A pie chart uses the relative area of a circle to indicate relative frequency instead of the relative area or frequency of a bar.
- A pie chart is rarely used because the size of relative area can be hard to judge.

Visualizing Small Numbers of Attributes

Percentile Plots

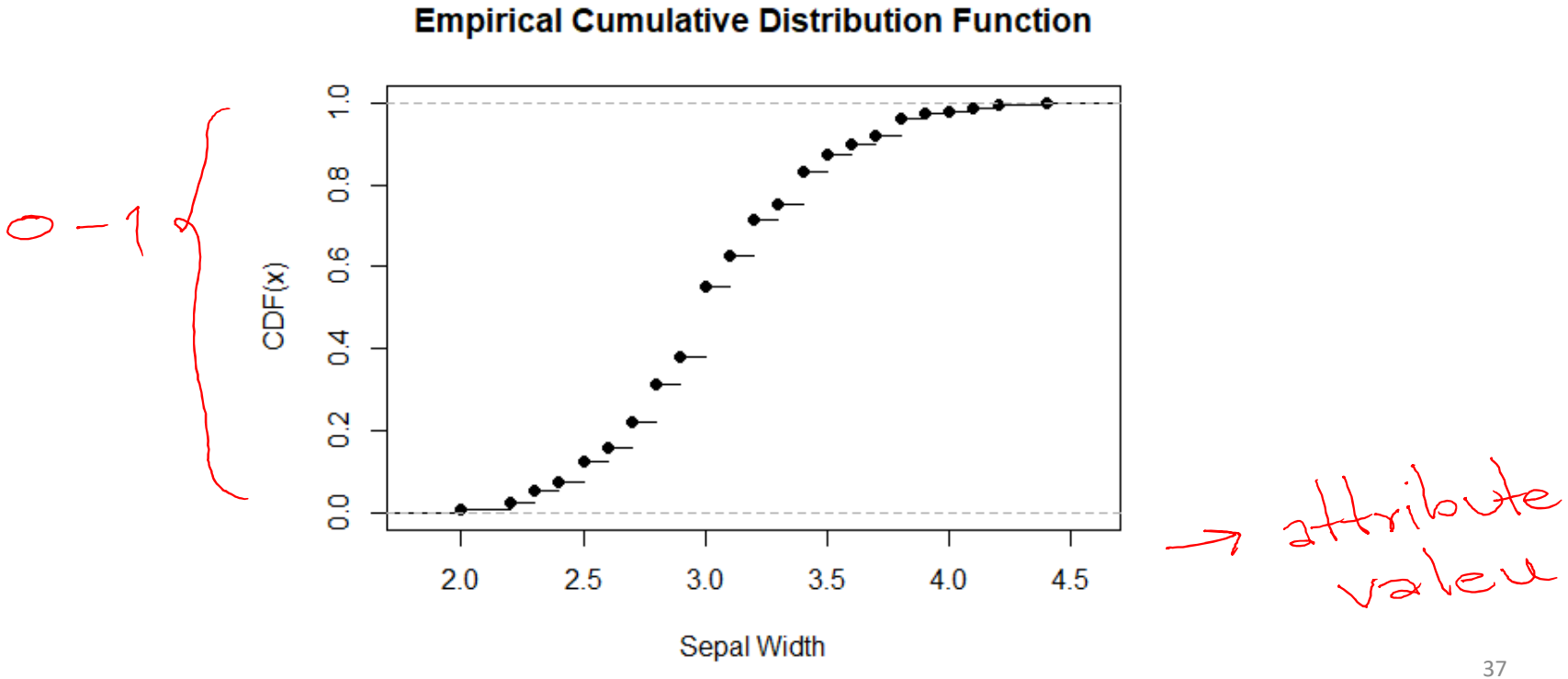
- Percentile plots (aka Quantile plots) display the percentiles (quantiles) of a set of values.
- The sample quantiles are plotted against the fraction of the sample they correspond to.



Visualizing Small Numbers of Attributes

Empirical Cumulative Distribution Functions

- Empirical Cumulative Distribution Functions is the step function associated with the empirical measure of a sample.



Visualizing Small Numbers of Attributes

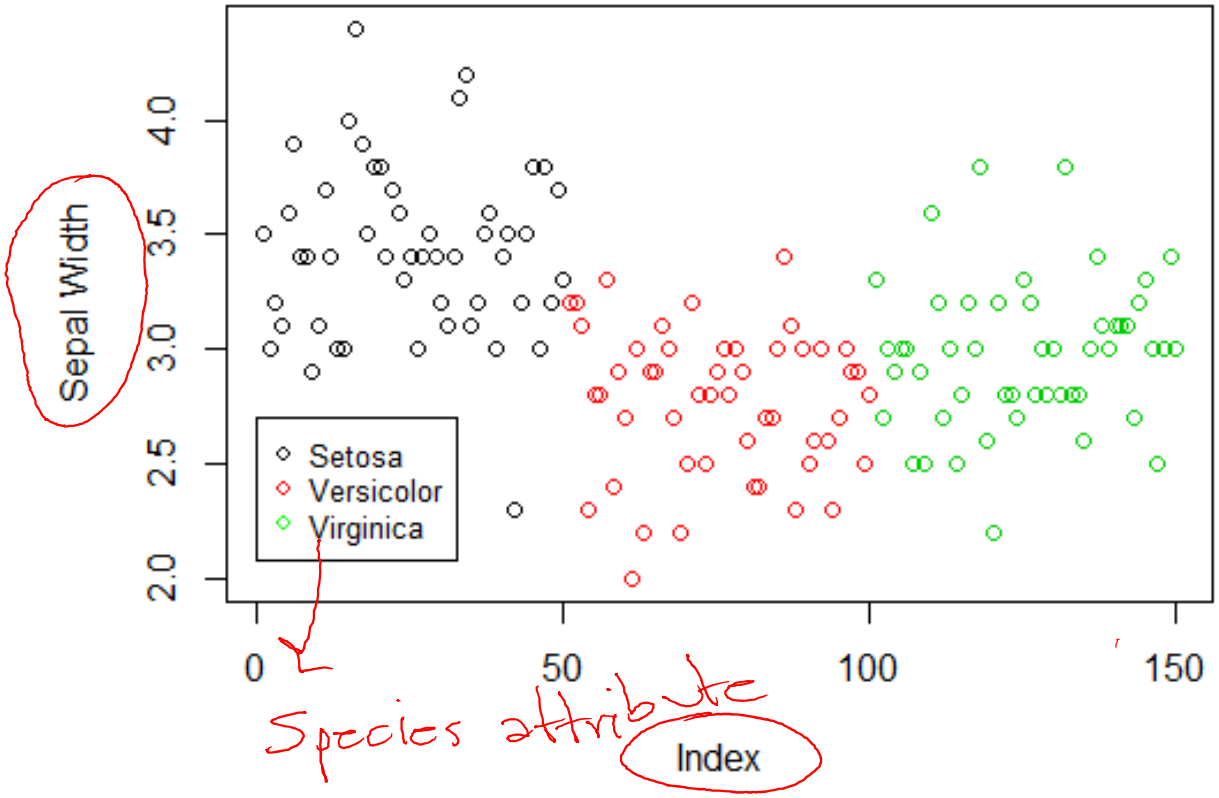
Scatter Plots

- Scatter plots can be used to illustrate linear correlation where each data object is plotted as a point in the plane using the values of the two attributes as x and y coordinates.

- Two main uses for scatter plots:
  - They graphically show the relationship between two attributes, such as degree of linear correlation, non-linear relationship
  - They can be used to investigate the degree to which two attributes separate the classes when class labels are available.

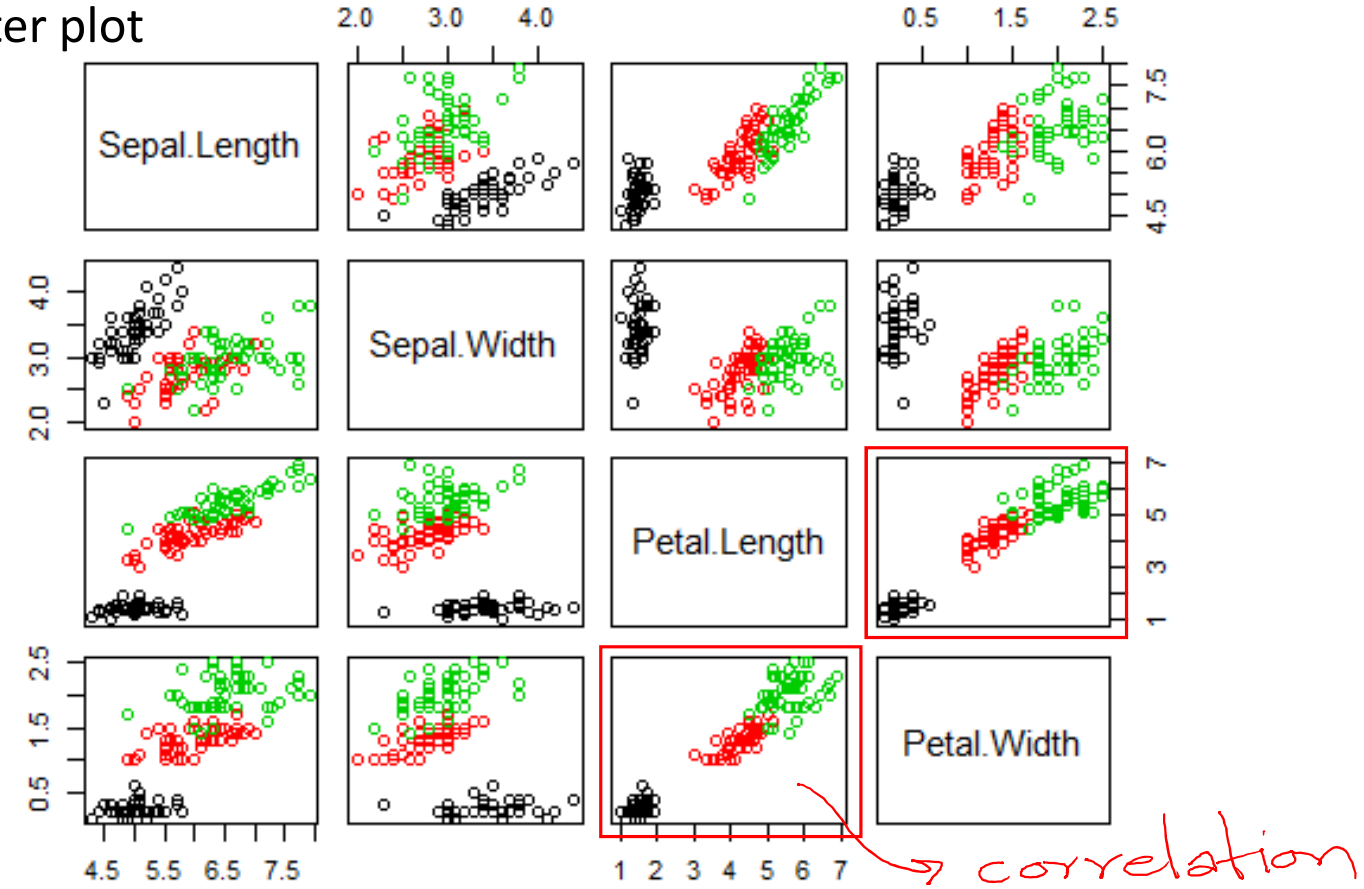
Visualizing Small Numbers of Attributes

2 attributes scatter plot



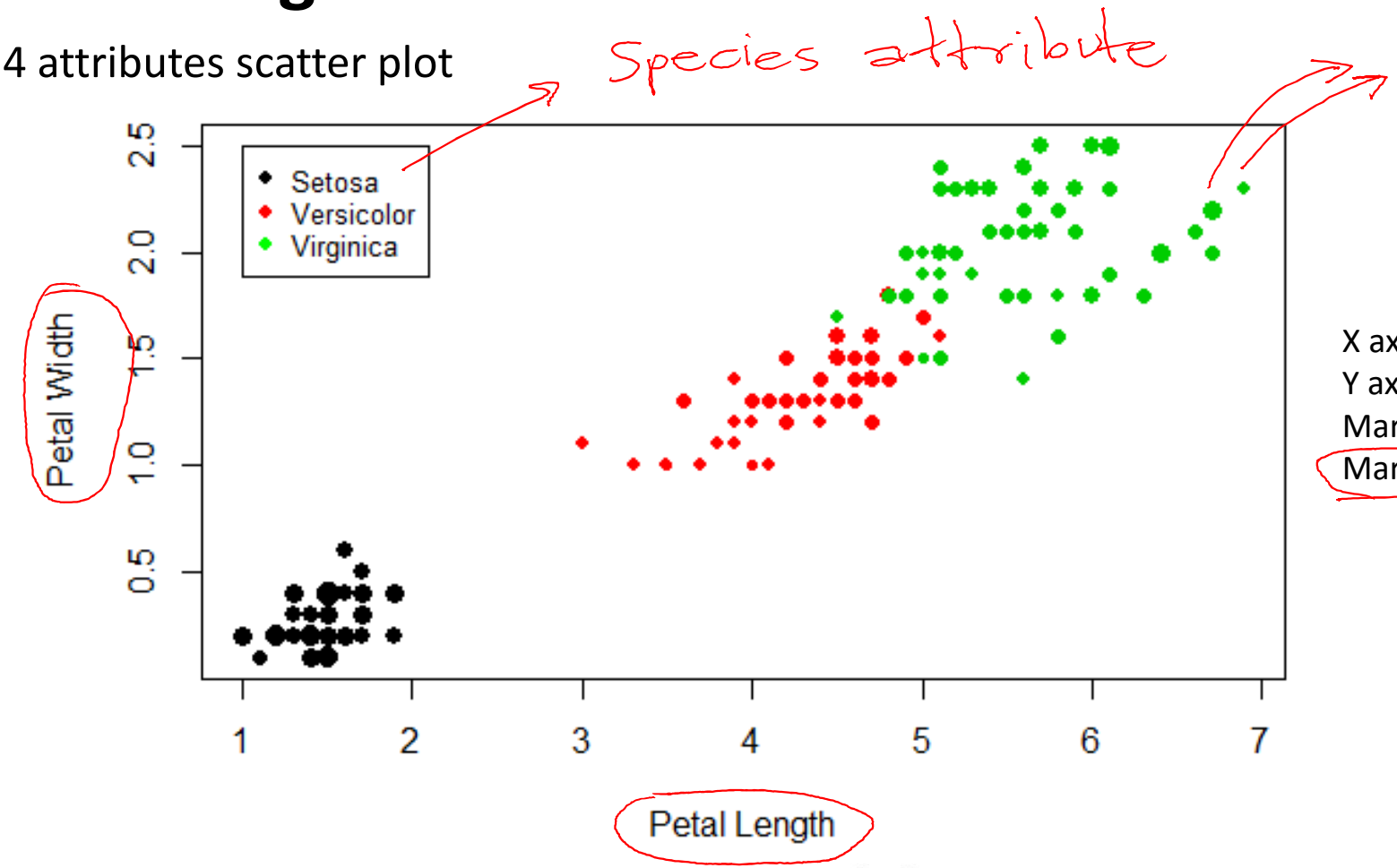
Visualizing Small Numbers of Attributes

3 attributes scatter plot



Visualizing Small Numbers of Attributes

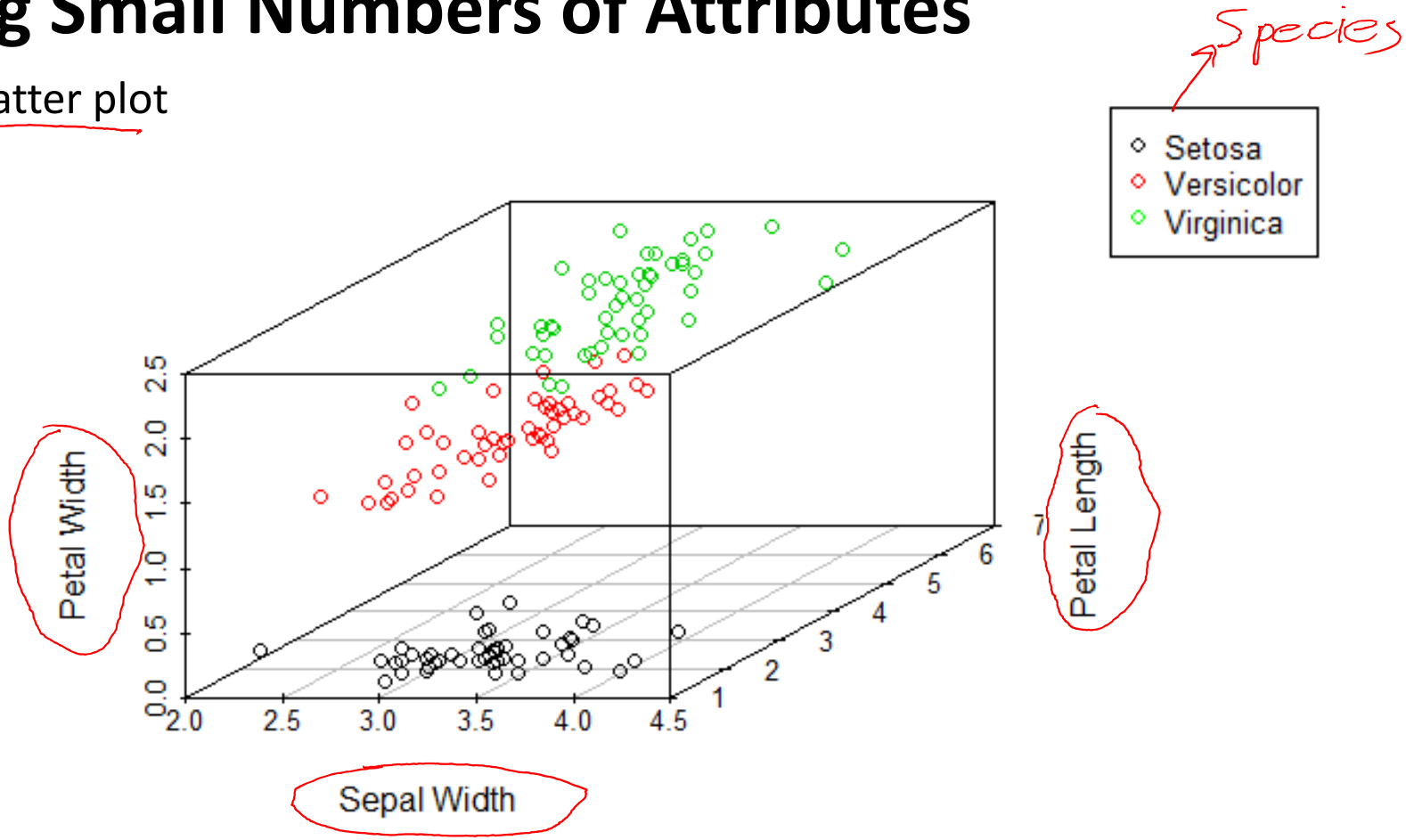
4 attributes scatter plot



X axis = Petal Length
Y axis = Petal Width
Marker color = Species
Marker size = Sepal Width

Visualizing Small Numbers of Attributes

4 attributes scatter plot



Visualizing Small Numbers of Attributes

Contour plots

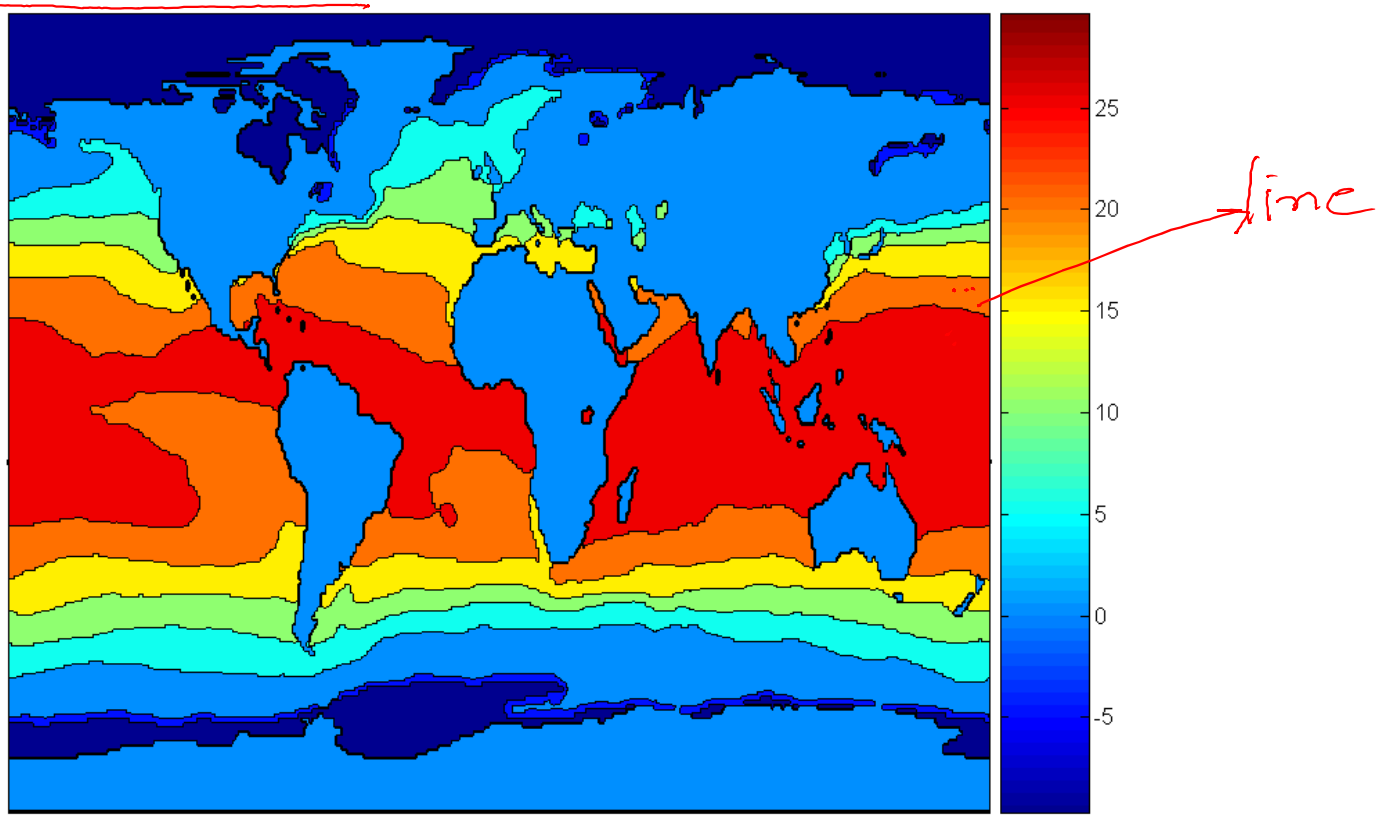
- Contour plot is useful for three-dimensional data where two attributes specify a position in a plane, such as latitude and longitude x and y coordinate, while the third has a continuous value, such as temperature or elevation.

- Contour plot breaks the plane into separate regions where the values of the third attribute (temperature, elevation) are roughly the same. These regions are separated by connecting points with equal values to form the contour lines which are considered as the boundaries.

- The most common example is contour maps of elevation, temperature, rainfall, air pressure.

Visualizing Small Numbers of Attributes

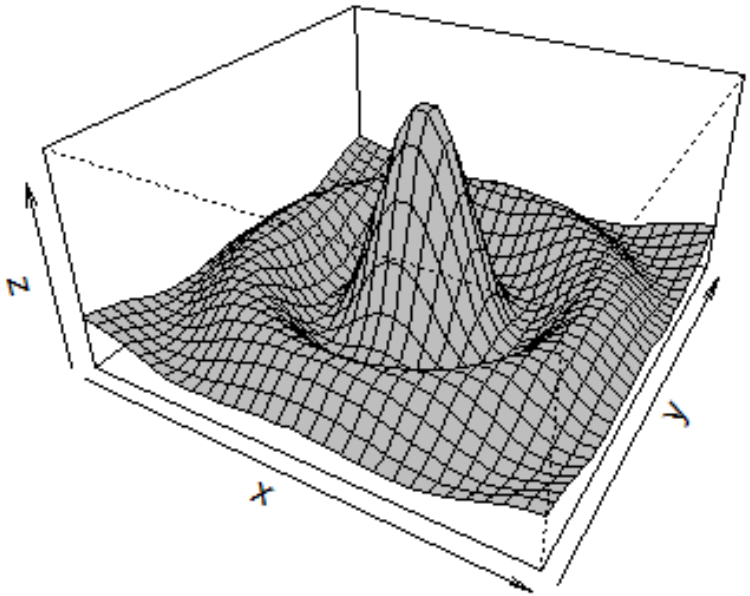
Example for Sea Surface Temperature (SST) for December 1998



Visualizing Small Numbers of Attributes

Surface Plots

- Surface plots are similar to contour plots where 2 attributes used to specify a position in a plane and the third attribute is used to indicate the height above the plane defined by the first two attributes.
- The limitation of surface plots is that a value of the third attribute must be defined for all combinations of values for the first two attributes.
- Surface plots are often used to describe the mathematical functions or physical surfaces that vary in a relatively smooth manner.



Topics

- Summary Statistics
- Visualization Techniques
  - Visualizing Small Numbers of Attributes
  - **Visualizing Higher-Dimensional Data**

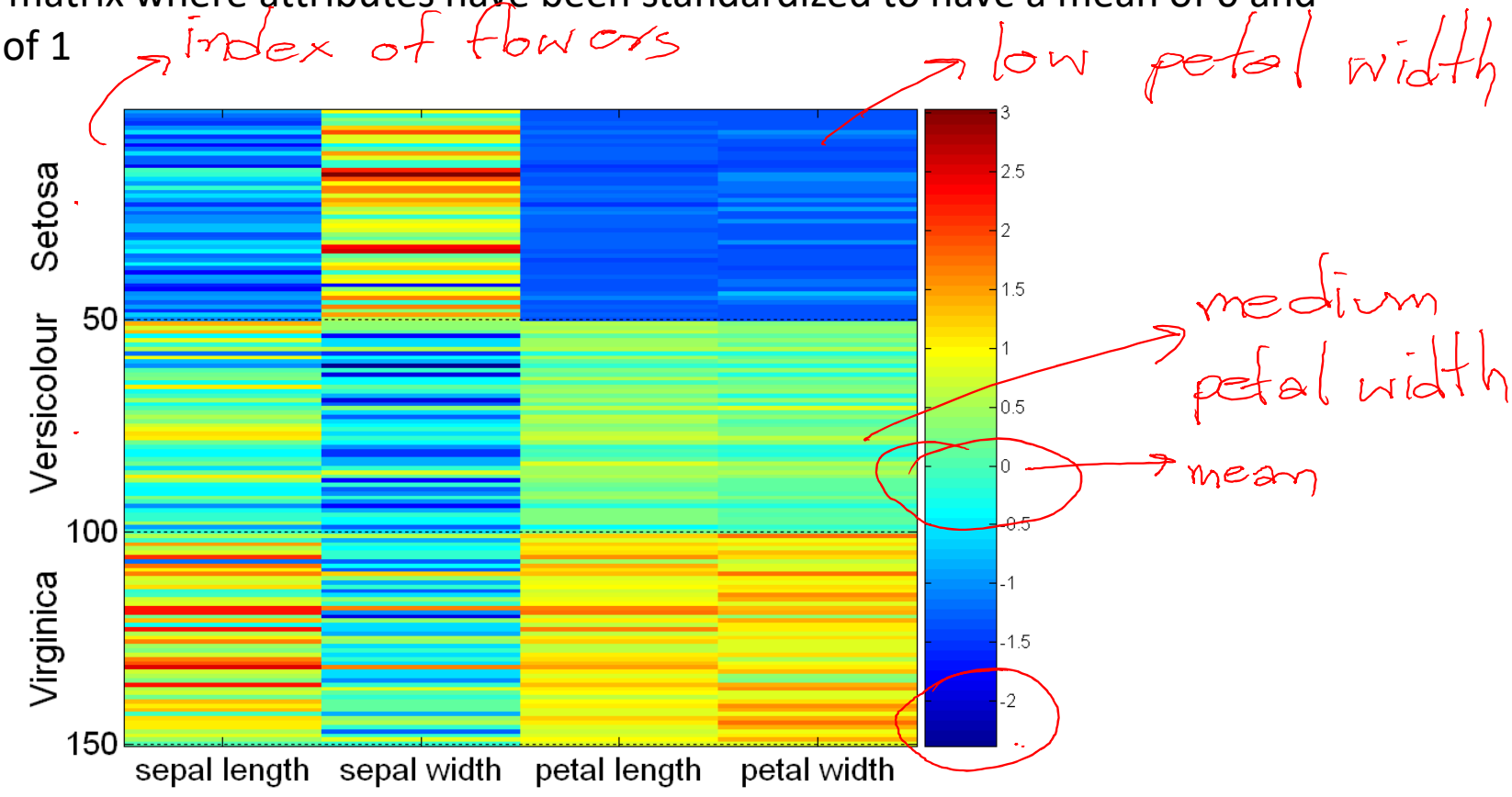
Visualizing Higher-Dimensional Data

Matrix Plots (Matrices)

- Matrix plots are used to visualize the data matrix of various attributes.
- It associates value of each entry in the data matrix with color and brightness.
- If class labels are known, it is useful to reorder the data matrix so that all objects of a class are together.
- If class labels are not known, various techniques (matrix reordering and seriation) can be used to rearrange the rows and columns of the similarity matrix so that groups of highly similar objects and attributes are together.
- Typically, the attributes are normalized to prevent the attribute with the largest values from visually dominating the plot.
- Plots of similarity or dissimilarity or correlation matrices can also be useful for visualizing the relationships between objects.

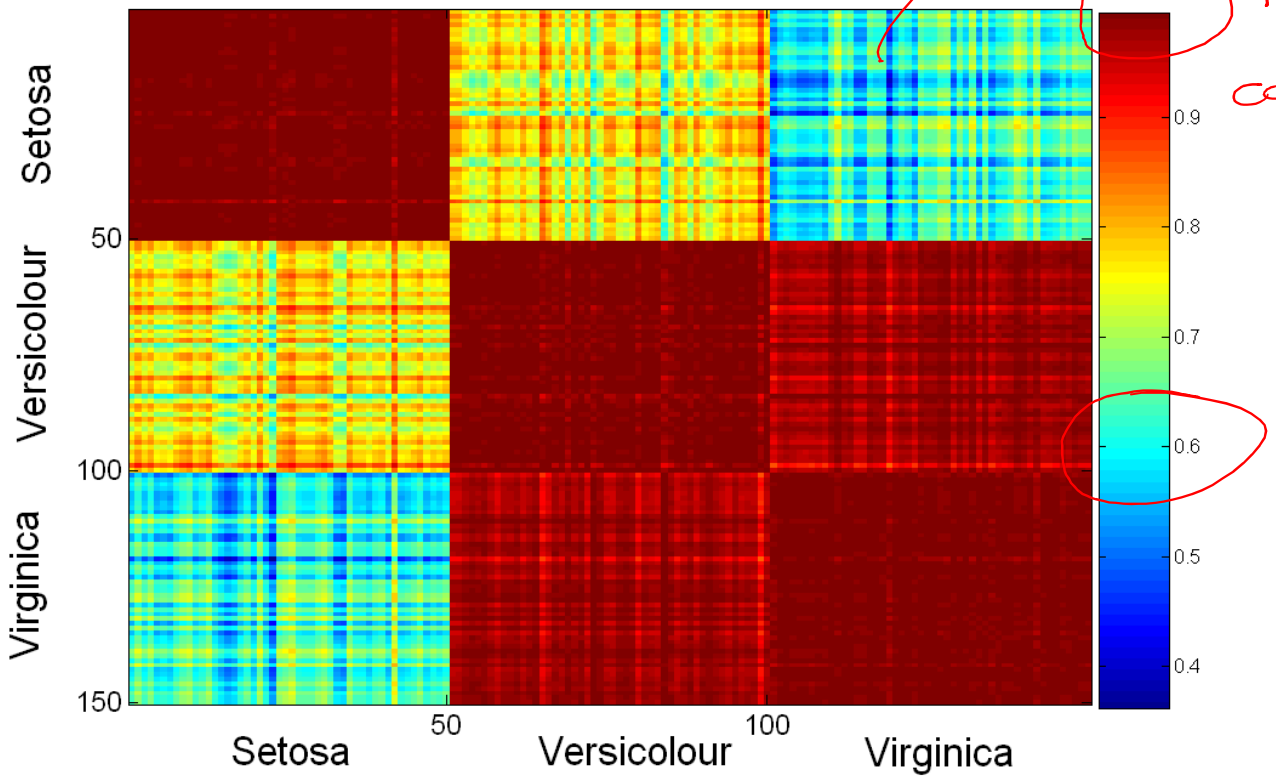
Visualizing Higher-Dimensional Data

Plot of the iris data matrix where attributes have been standardized to have a mean of 0 and standard deviation of 1



Visualizing Higher-Dimensional Data

Plot of the iris correlation matrix



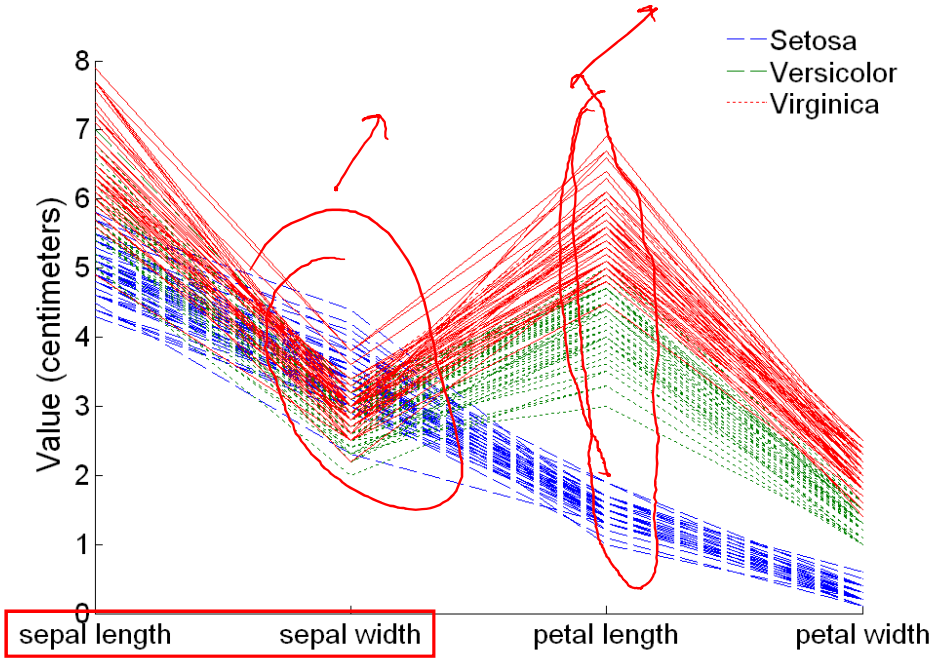
Note: Versicolor and Virginica are more similar to one another than to Setosa.

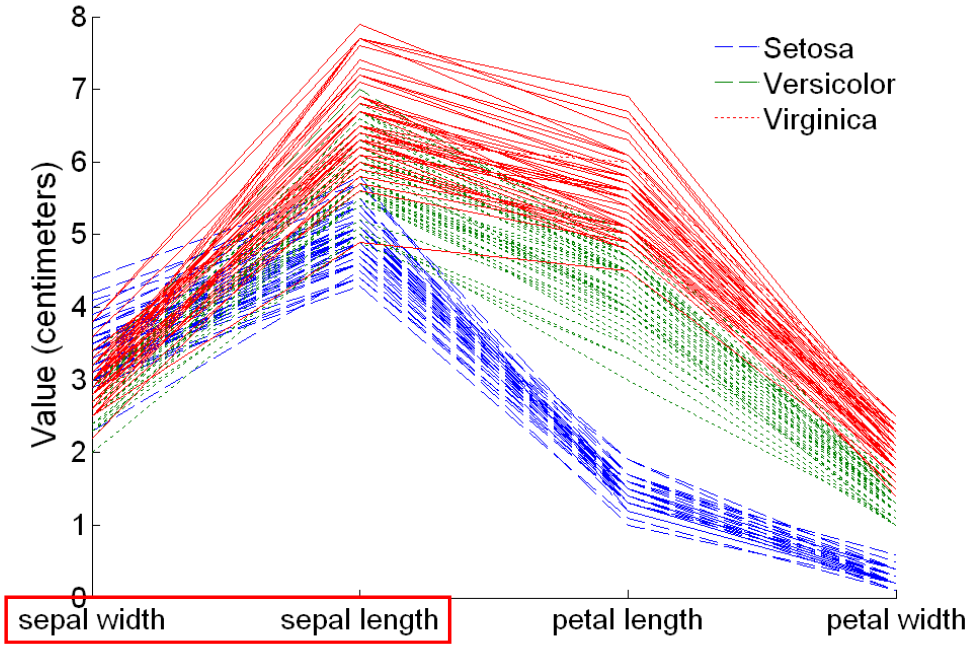
Visualizing Higher-Dimensional Data

Parallel Coordinates

- Parallel coordinates have one coordinate axis for each attribute, but the different axes are parallel to one other instead of perpendicular.
- A data object is represented as a line instead of as a point.
- The value of each attribute of an object is mapped to a point on the coordinate axis associated with that attribute, and these points are then connected to form the line that represents the data object.
- The lines representing a distinct class of objects often group together, at least for some attributes.
- The drawback of this plot is that the detection of patterns in such a plot may depend on the order.

Visualizing Higher-Dimensional Data



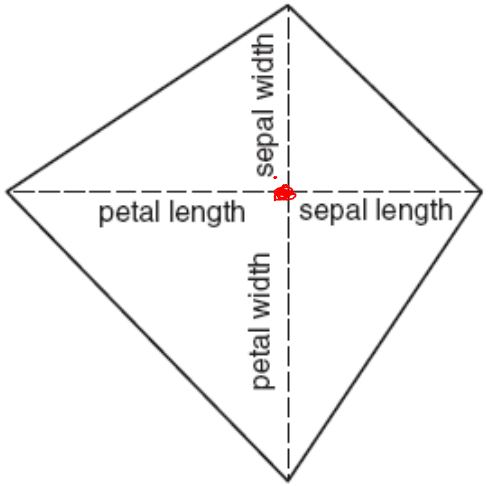


Reordered attributes

Visualizing Higher-Dimensional Data

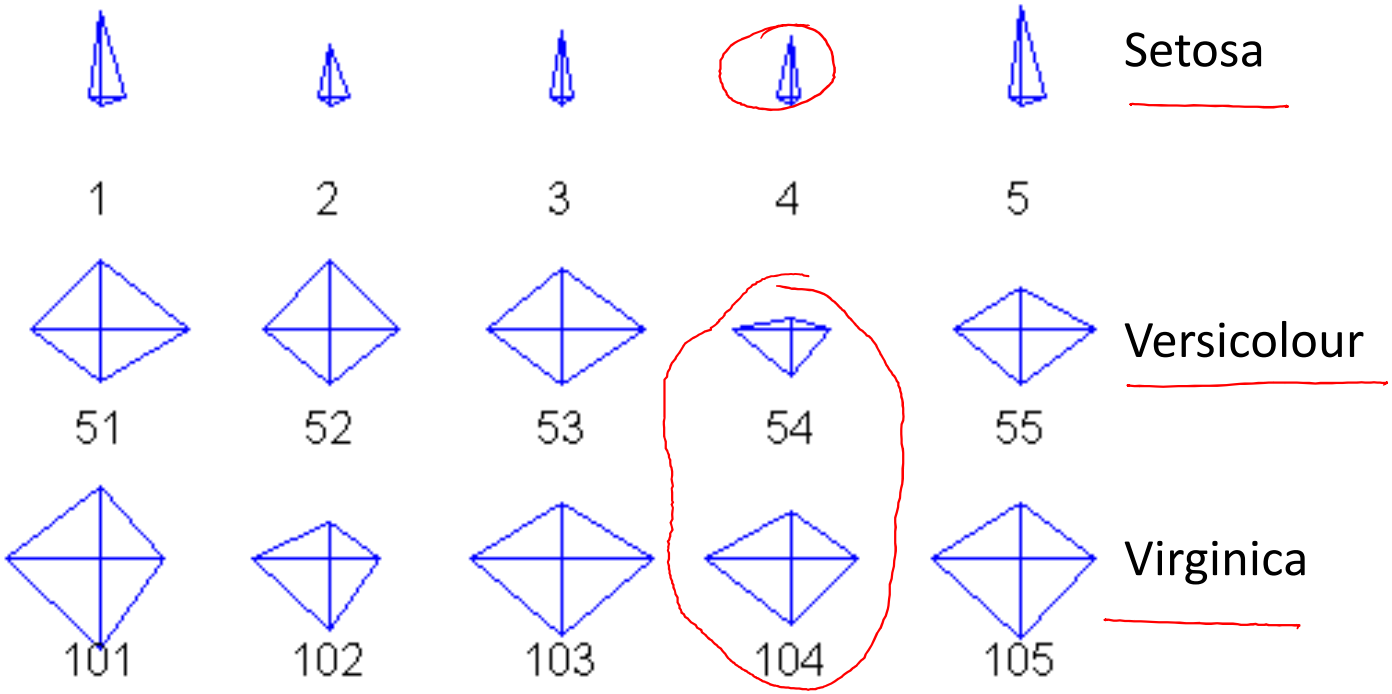
Star Coordinates

- Star coordinates is similar approach to parallel coordinates, but axes radiate from a central point.
- Typically, all the attribute values are mapped to the range [0,1] by converting each attribute value of the object to a fraction that represents its distance between the minimum and maximum value of the attribute.
- These points are connects by line segments to form a polygon.
- The size and shape of this polygon gives a visual description of the attribute values of the object.



Visualizing Higher-Dimensional Data

Star Plots for Iris Data



Visualizing Higher-Dimensional Data

Chernoff Faces

- This approach associates each attribute with a characteristic of a face, and the attribute value is used to determine the way that the facial feature is expressed
- The following table is the four numerical attributes of Iris data set face features.

Data Feature	Facial Feature
Sepal length	Size of face
Sepal width	Forehead/jaw relative arc length
Petal length	Shape of forehead
Petal width	Shape of jaw

Visualizing Higher-Dimensional Data

Chernoff Faces for Iris Data