STATISTICAL METHODS IN DATA MINING

DR. ALPER VAHAPLAR

Exploring Data

- ✓Descriptive Statistics
- ✓Data Visualization
- ✓Graphs and Tables

References:

- ✓*Han, J. , Kamber, M., Pei, J., (2011). Data Mining: Concepts and Techniques.*
- ✓*Larose, Daniel T. (2005). Discovering Knowledge In Data – An Introduction to Data Mining.*
- ✓*Tan, P., Steinbach, M., Kumar, v. (2006) Introduction to Data Mining.*
- ✓*Bramer, M., (2007) Principles of Data Mining.*
- ✓*Birant, D. Lecture Notes (2012).*

Exploring Data

- ✓ Data understanding,
- ✓ **A preliminary exploration of the data to better understand its characteristics.**
- ✓ Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools
- ✓ Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey

Exploring Data

- ✓ In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques

- ✓ The Iris Plant data set.

  - http://alpervahaplar.com - iris.xls

- ✓ **Attribute Information:**

- ✓ 1. sepal length in cm
  2. sepal width in cm
  3. petal length in cm
  4. petal width in cm
  5. class:
  -- Iris Setosa
  -- Iris Versicolour
  -- Iris Virginica

Iris Setosa

Iris Virginica

Iris Versicolor

Exploring Data

- ✓ Summary Statistics
  - Frequencies and Mode
  - Quartiles, Percentiles
  - Measures of Location (Central Tendency)
    - Mean
    - Median
  - Measures of Spread (Dispersion)
    - Range,
    - Standard Deviation,
    - Variance
  - Multivariate Summary Statistics

Summary Statistics

- Frequency and Mode
- The *frequency* of an attribute value is the percentage of time the value occurs in the data set.

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

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

Summary Statistics

- ✓ 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.
- ✓ By tradition, $\min(x) = x_{0\%}$, $\max(x) = x_{100\%}$
- ✓ For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.
- ✓ Quartiles (4)
- ✓ Quintiles (5)
- ✓ Deciles (10)

Summary Statistics

- ✓ Measures of Location (Central Tendency)
  - ***Mean***

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

  - ***Median***

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

Summary Statistics

- ✓ The mean is the most common measure of the location of a set of points.
- ✓ However, the mean is very sensitive to outliers.
- ✓ Trimmed mean
  - (a percentage *p* is specified, top and bottom (*p*/2)% of the data is thrown out, mean is calculated)
- ✓ Mean → *p=0%*, Median → *p=100%*
- ✓ *mean – mode = 3 x (mean – median)* (unimodal and skewed)
- ✓ Types of Mean
  - Arithmetic mean,
  - Weighted mean,
  - Trimmed mean,
  - Geometric mean,
  - Harmonic mean

Summary Statistics

- ✓ Means and medians for Iris data (values in cm.)

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
tr. mean (p=20%)	5.79	3.02	3.72	1.12

Summary Statistics

- ✓ Measures of Dispersion
- ✓ Range = max(x) - min(x)
- ✓ Variance, Standard Deviation
- ✓ InterQuartile Range (IQR)
- ✓ Absolute Average Deviation (AAD)
- ✓ Median Absolute Deviation (MAD)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

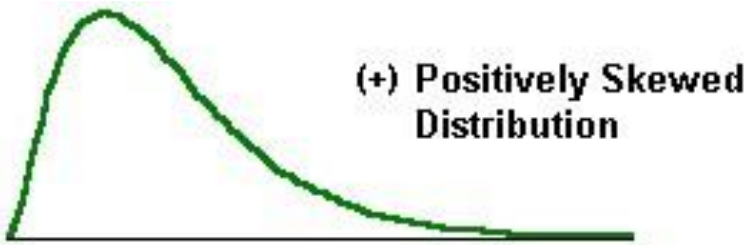
$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

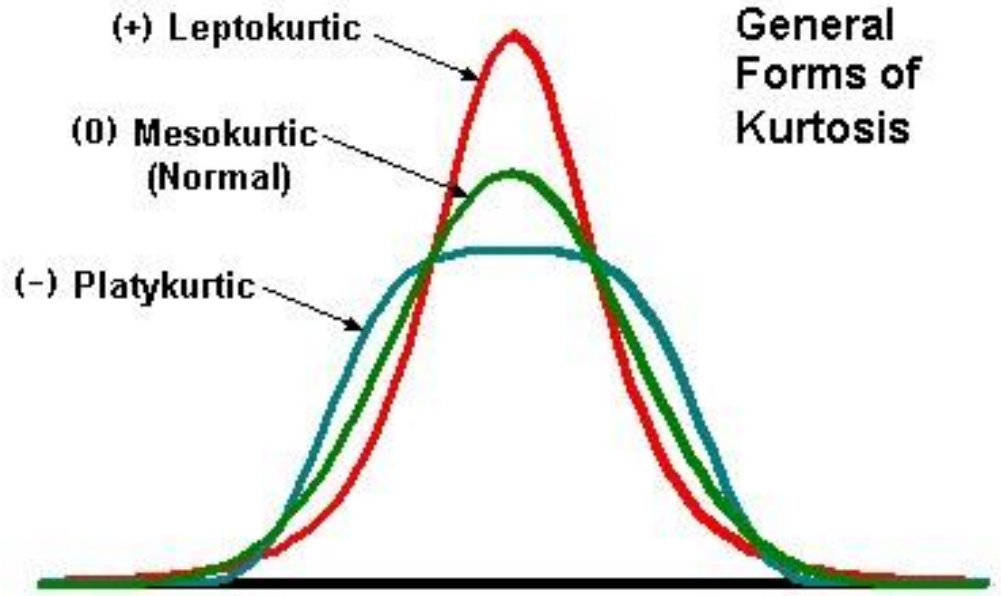
$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

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

Summary Statistics

- ✓ Skewness and Kurtosis





$$Coef.of\ Skewness = \frac{\mu_3}{\sigma^3} = \frac{E(X-\mu)^3}{\sigma^3}$$

$$Coef.of\ Kurtosis = \frac{\mu_4}{\sigma^4} = \frac{E(X-\mu)^4}{\sigma^4}$$

Summary Statistics

- ✓ Iris Data

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
tr. mean (p=20%)	5.79	3.02	3.72	1.12
range	3.6	2.4	5.9	2.4
std	0.8	0.4	1.8	0.8
IQR	1.3	0.5	3.5	1.5
AAD	0.7	0.3	1.6	0.6
MAD	0.7	0.3	1.2	0.7
Skewness	0.31	0.33	-0.27	-0.10
Kurtosis	-0.55	0.29	-1.40	-1.34

Summary Statistics

- ✓ 5 Number Summary
  1. Minimum
  2. First Quartile (Q1)
  3. Median
  4. Third Quartile (Q3)
  5. Maximum

Summary Statistics

- ✓ Multivariate Summary Statistics
- ✓ Covariance

$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- ✓ Correlation

$$r_{xy} = \frac{s_{x,y}}{s_x s_y}$$

Visualization

- ✓ the conversion of data into a ***visual*** or ***tabular*** format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
- ✓ The Goal:
  - interpretation of the visualized information by a person,
  - formation of a mental model of the information.
- ✓ 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

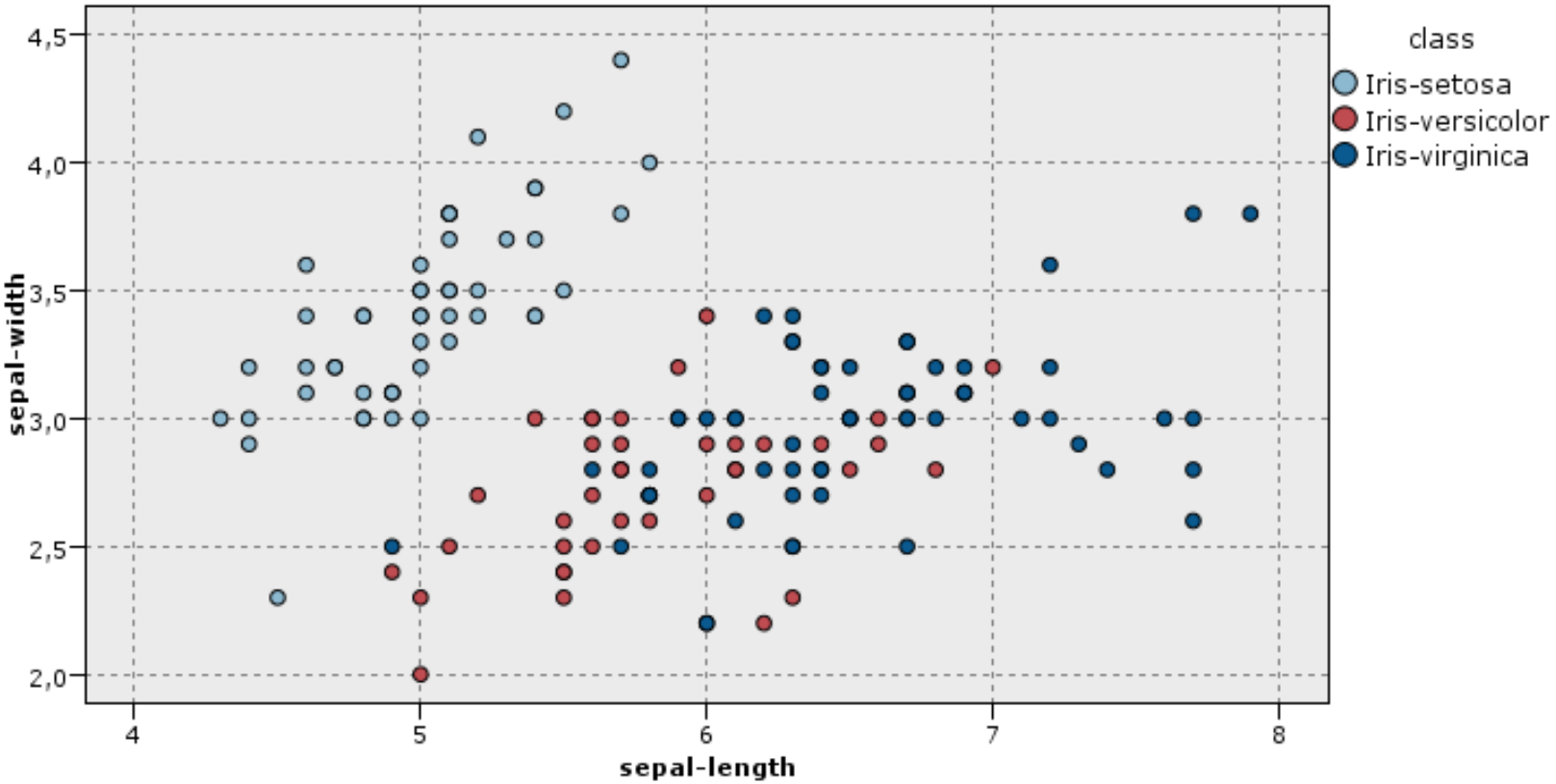
sepal-length	sepal-width	petal-length	petal-width	class
5.500	2.500	4.000	1.300	Iris-versicolor
5.500	2.600	4.400	1.200	Iris-versicolor
6.100	3.000	4.600	1.400	Iris-versicolor
5.800	2.600	4.000	1.200	Iris-versicolor
5.000	2.300	3.300	1.000	Iris-versicolor
5.600	2.700	4.200	1.300	Iris-versicolor
5.700	3.000	4.200	1.200	Iris-versicolor
5.700	2.900	4.200	1.300	Iris-versicolor
6.200	2.900	4.300	1.300	Iris-versicolor
5.100	2.500	3.000	1.100	Iris-versicolor
5.700	2.800	4.100	1.300	Iris-versicolor
6.300	3.300	6.000	2.500	Iris-virginica
5.800	2.700	5.100	1.900	Iris-virginica
7.100	3.000	5.900	2.100	Iris-virginica
6.300	2.900	5.600	1.800	Iris-virginica
6.500	3.000	5.800	2.200	Iris-virginica
7.600	3.000	6.600	2.100	Iris-virginica
4.900	2.500	4.500	1.700	Iris-virginica
7.300	2.900	6.300	1.800	Iris-virginica

Visualization

- ✓ General Concepts:
  - Representation
    - Is the mapping of 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.
  - Arrangement
    - Is the placement of visual elements within a display.
    - Can make a large difference in how easy it is to understand the data.
  - Selection
    - Is the elimination or the de-emphasis of certain objects and attributes.

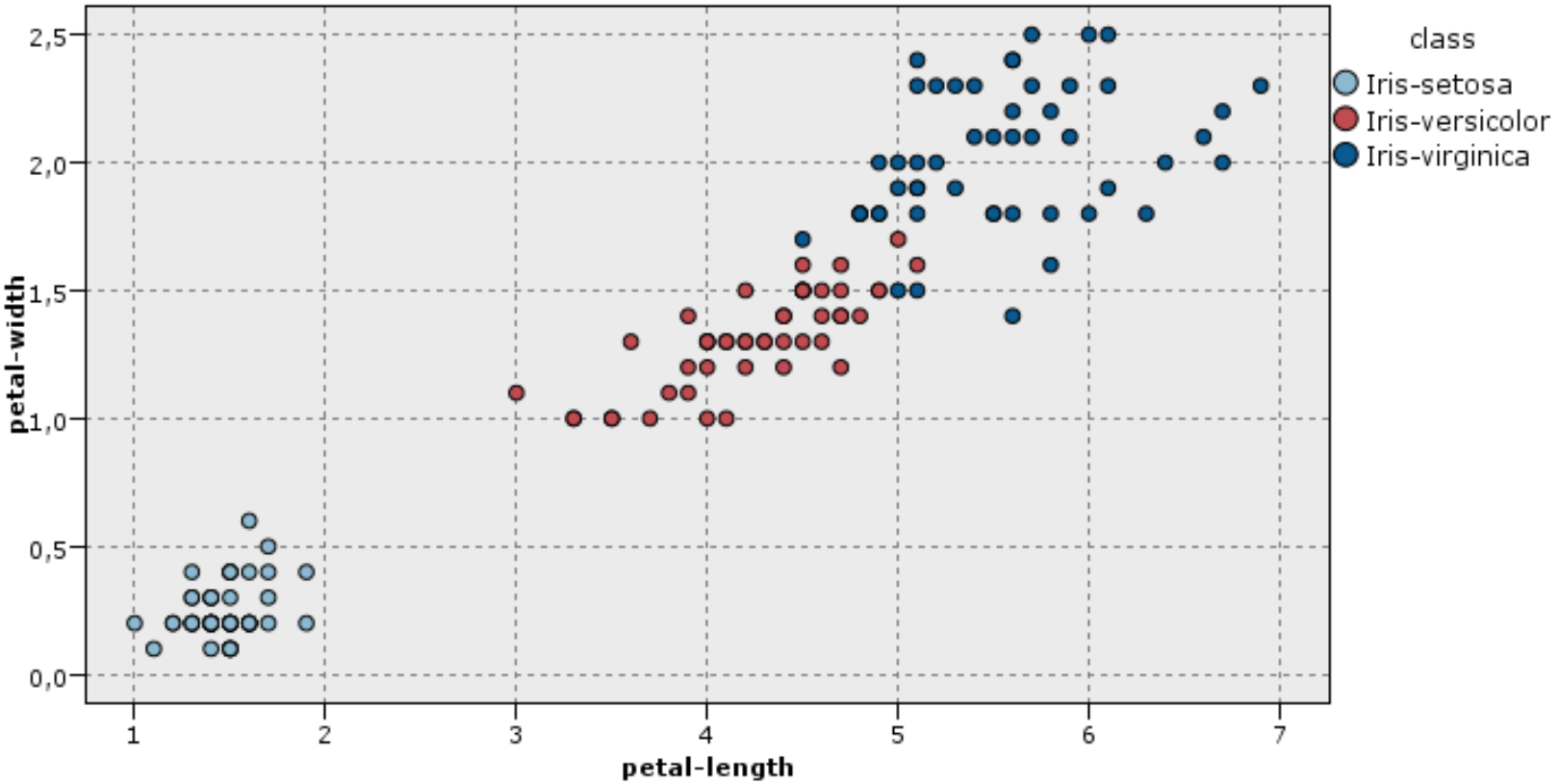
Visualization

- ✓ Representing



Visualization

- ✓ Arrangement



Visualization

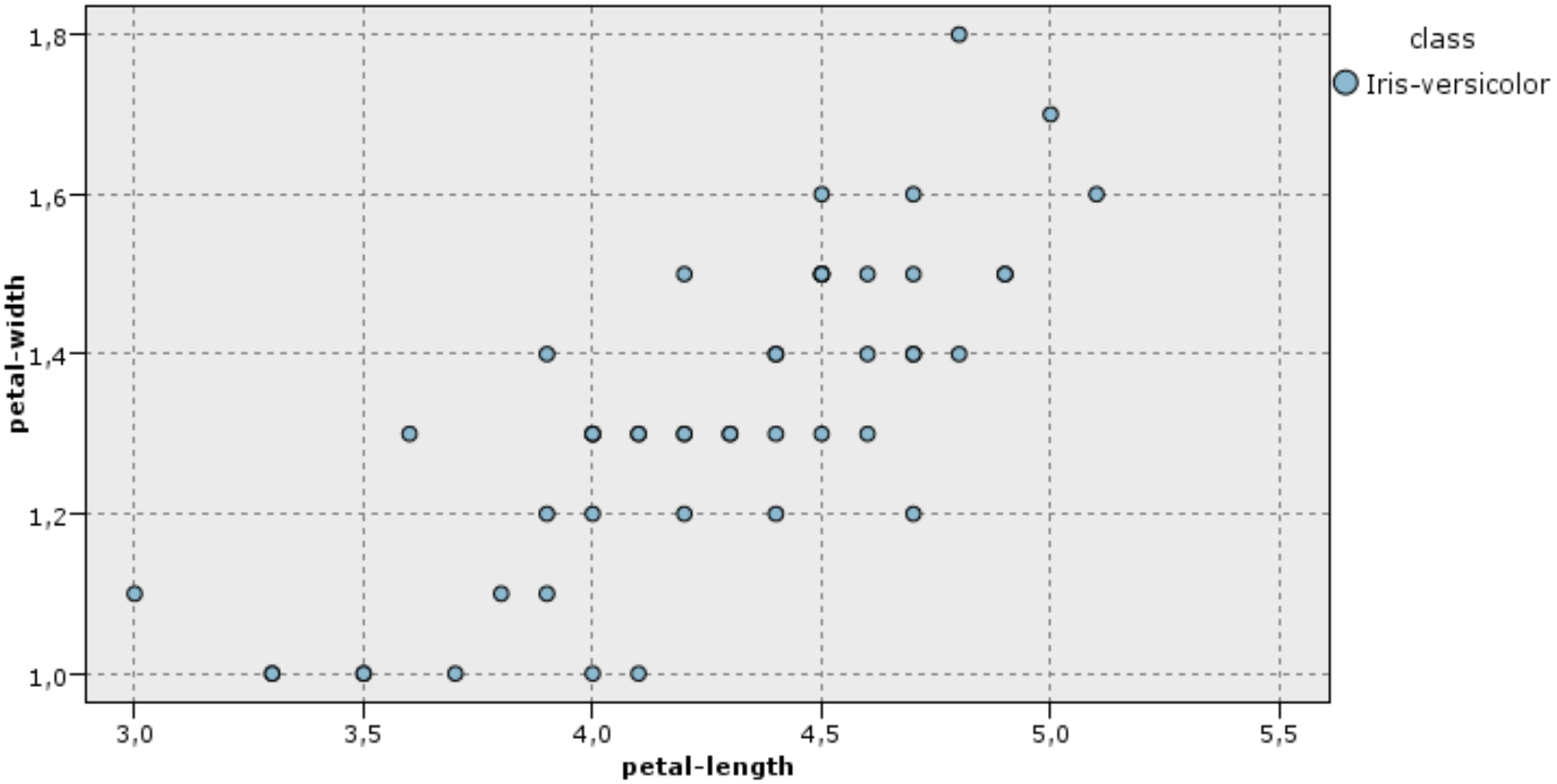
- ✓ Arrangement

	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



Visualization

- ✓ Techniques
  - Stem and Leaf Graphs
  - Bar Charts, Pie Charts
  - Histograms
  - Box Plots
  - Scatter Plots
  - Contour Plots
  - Surface Plots
  - Star Graph
  - Chernoff Faces

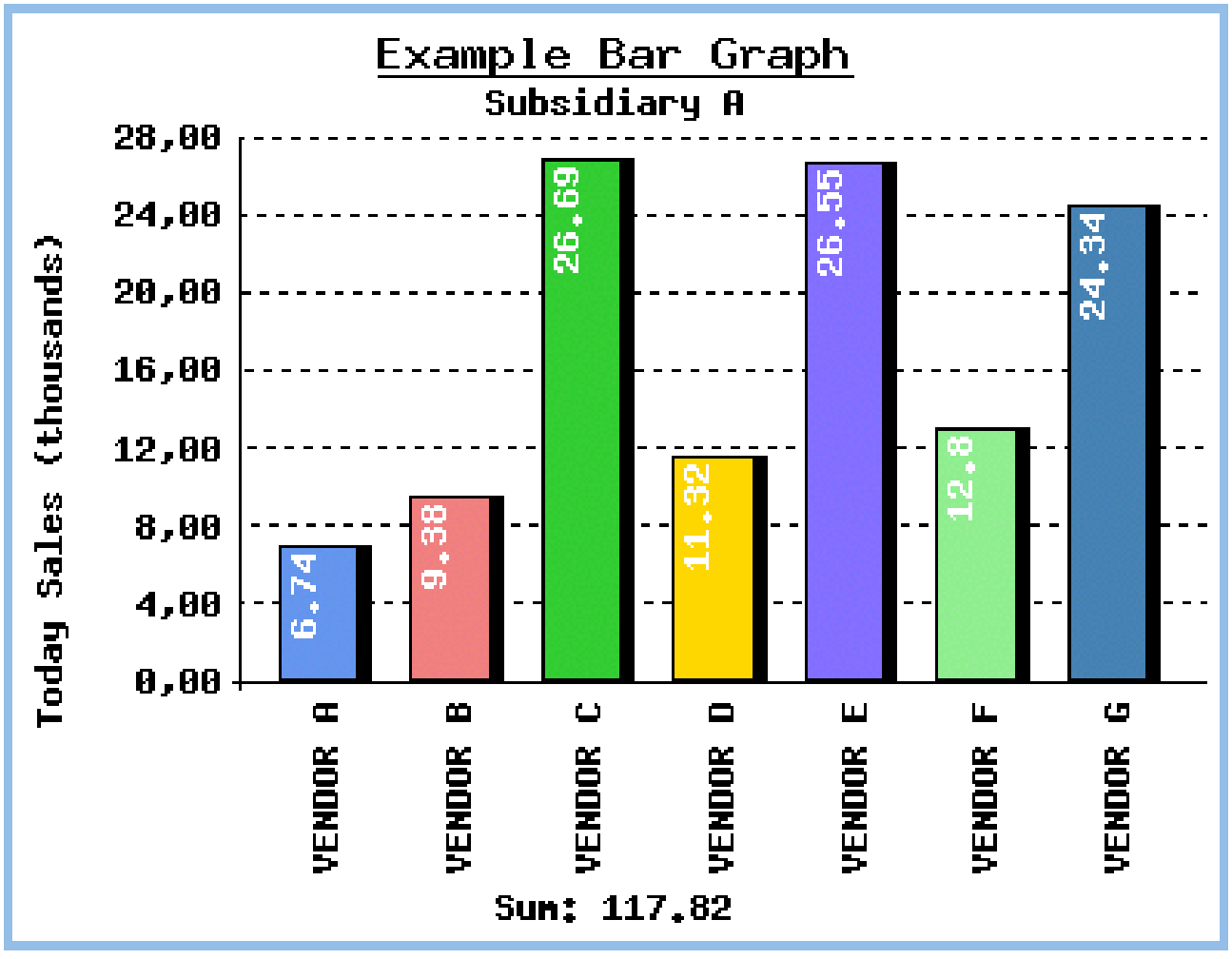
Visualization

- ✓ Stem and Leaf Plot

```
Variable: sepall
4  :  3444
4  :  566667788888999999
5  :  000000000011111111122223444444
5  :  55555556666667777777788888888999
6  :  000000111111222233333333334444444
6  :  555556677777778889999
7  :  0122234
7  :  677779
```

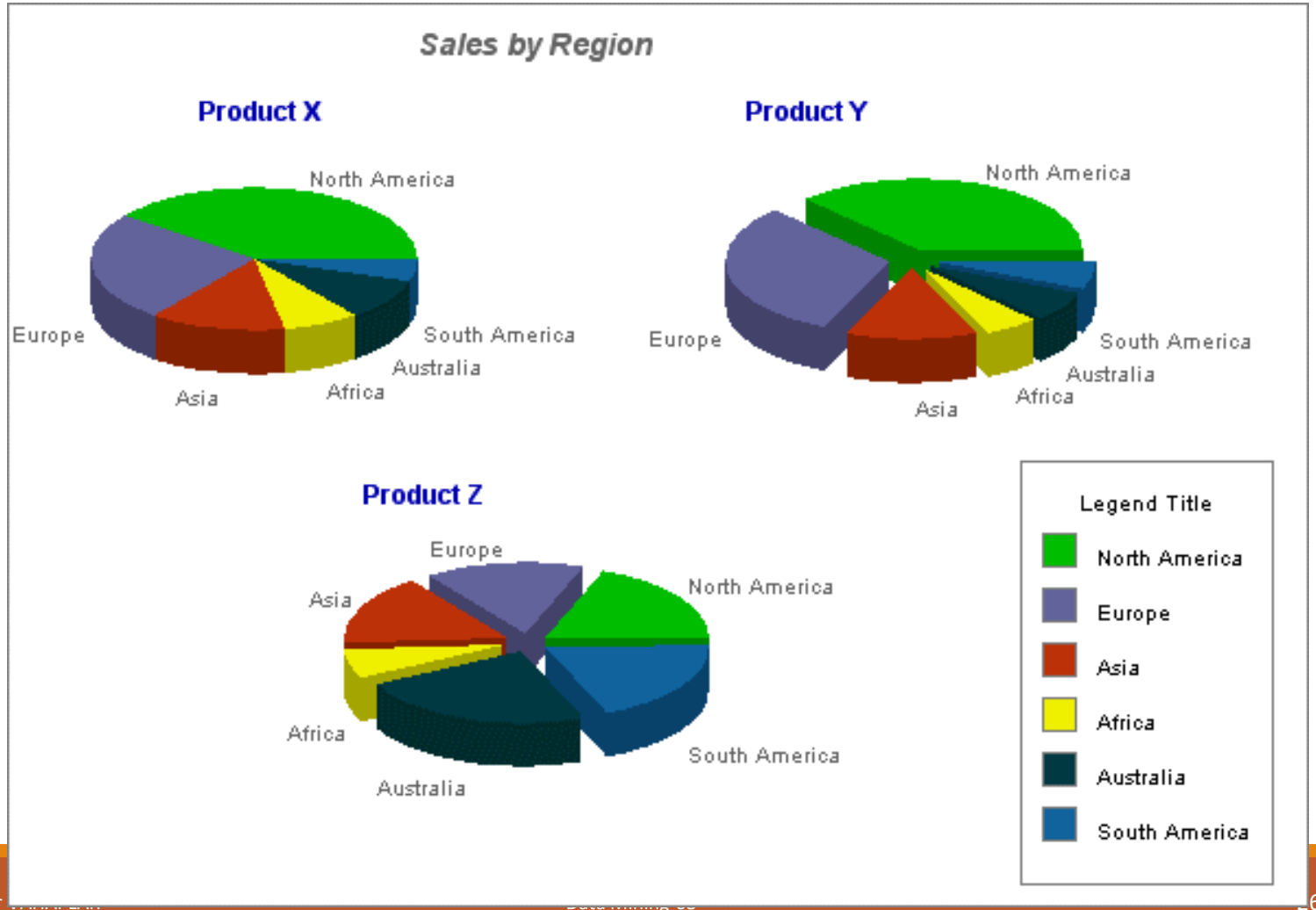
Visualization

- ✓ Bar charts and Pie charts



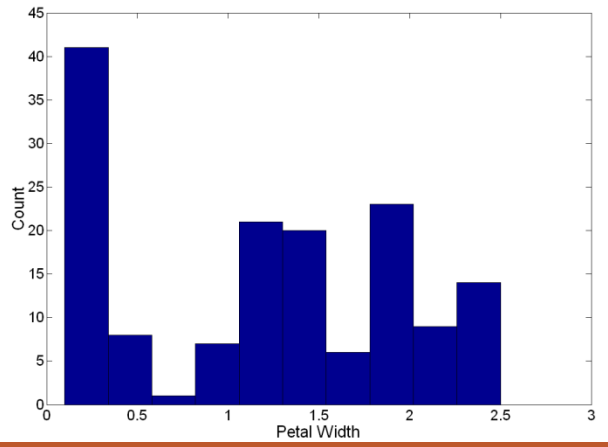
Visualization

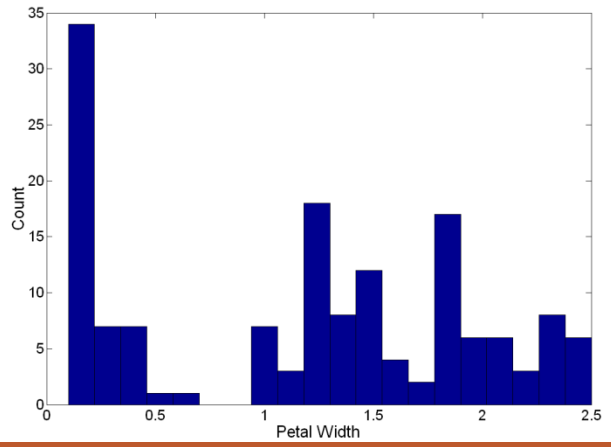
- ✓ Bar charts and Pie charts



Visualization

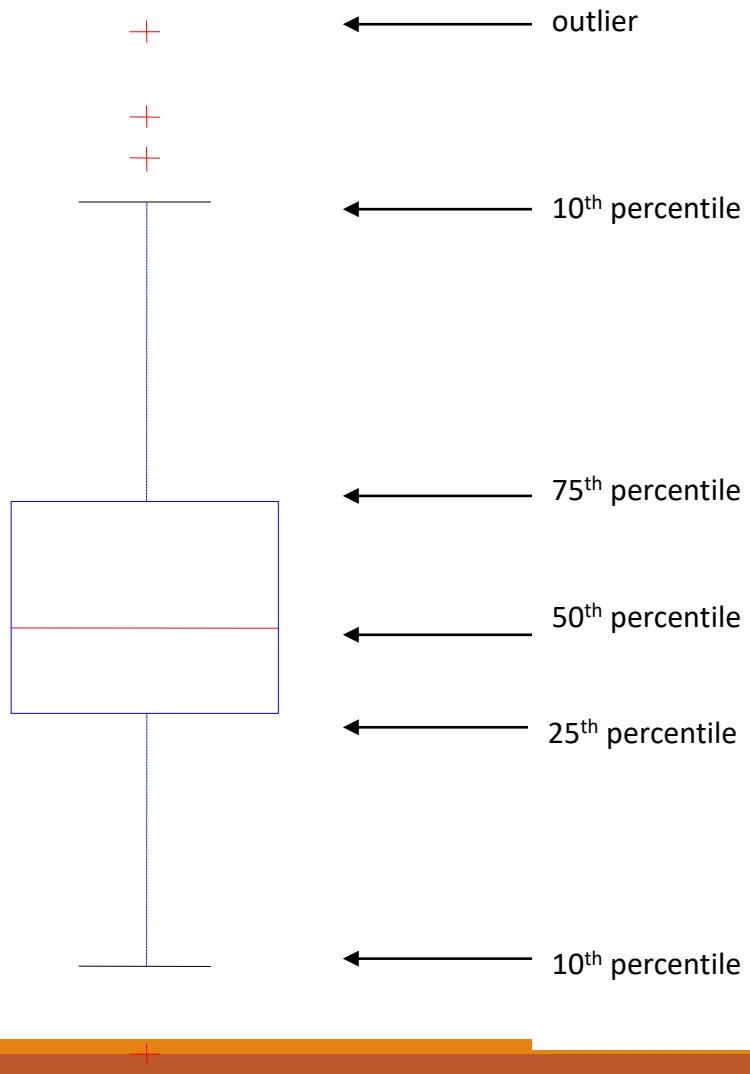
- ✓ Histograms
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- ✓ Example: Petal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





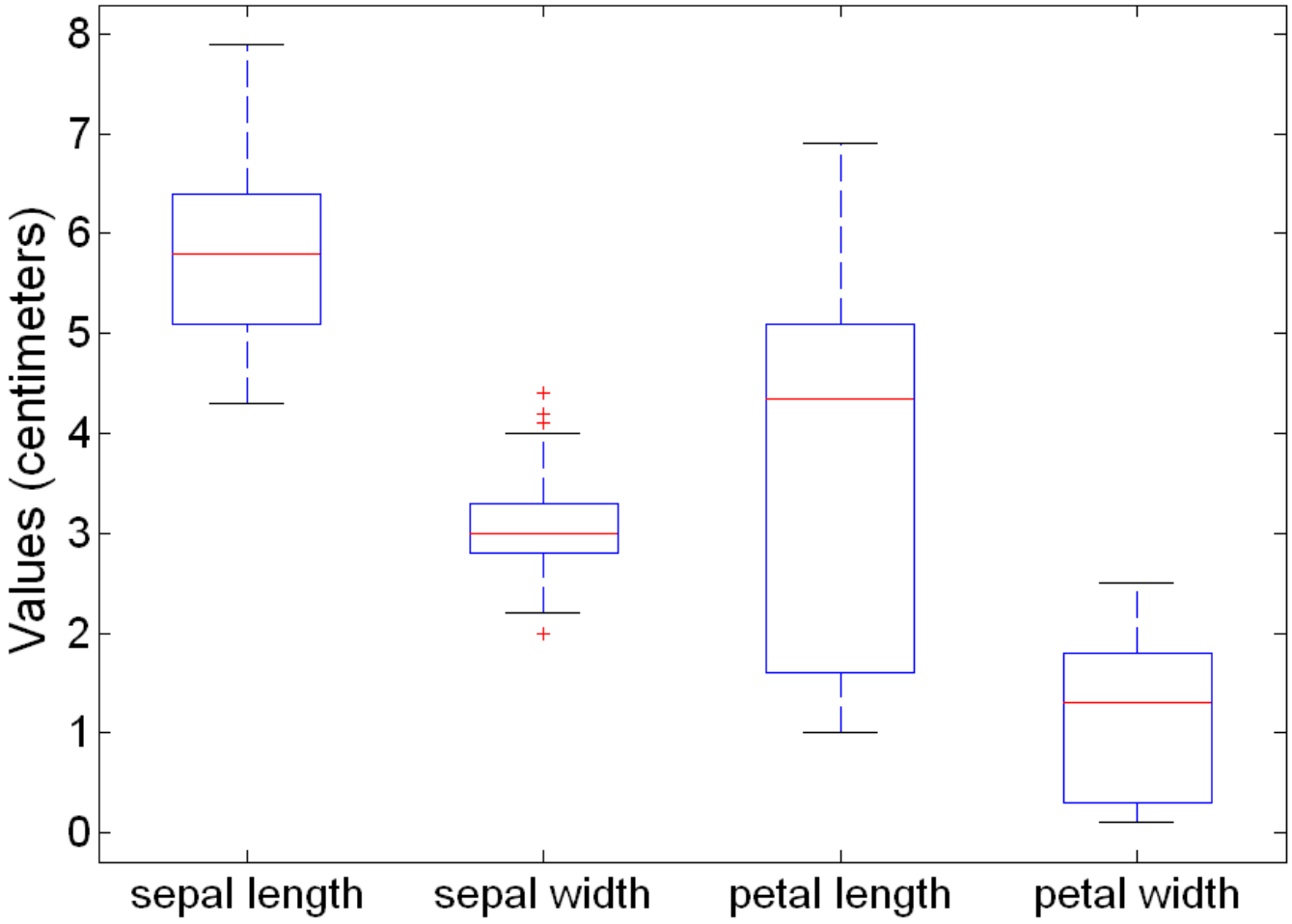
Visualization

- ✓ Box Plots



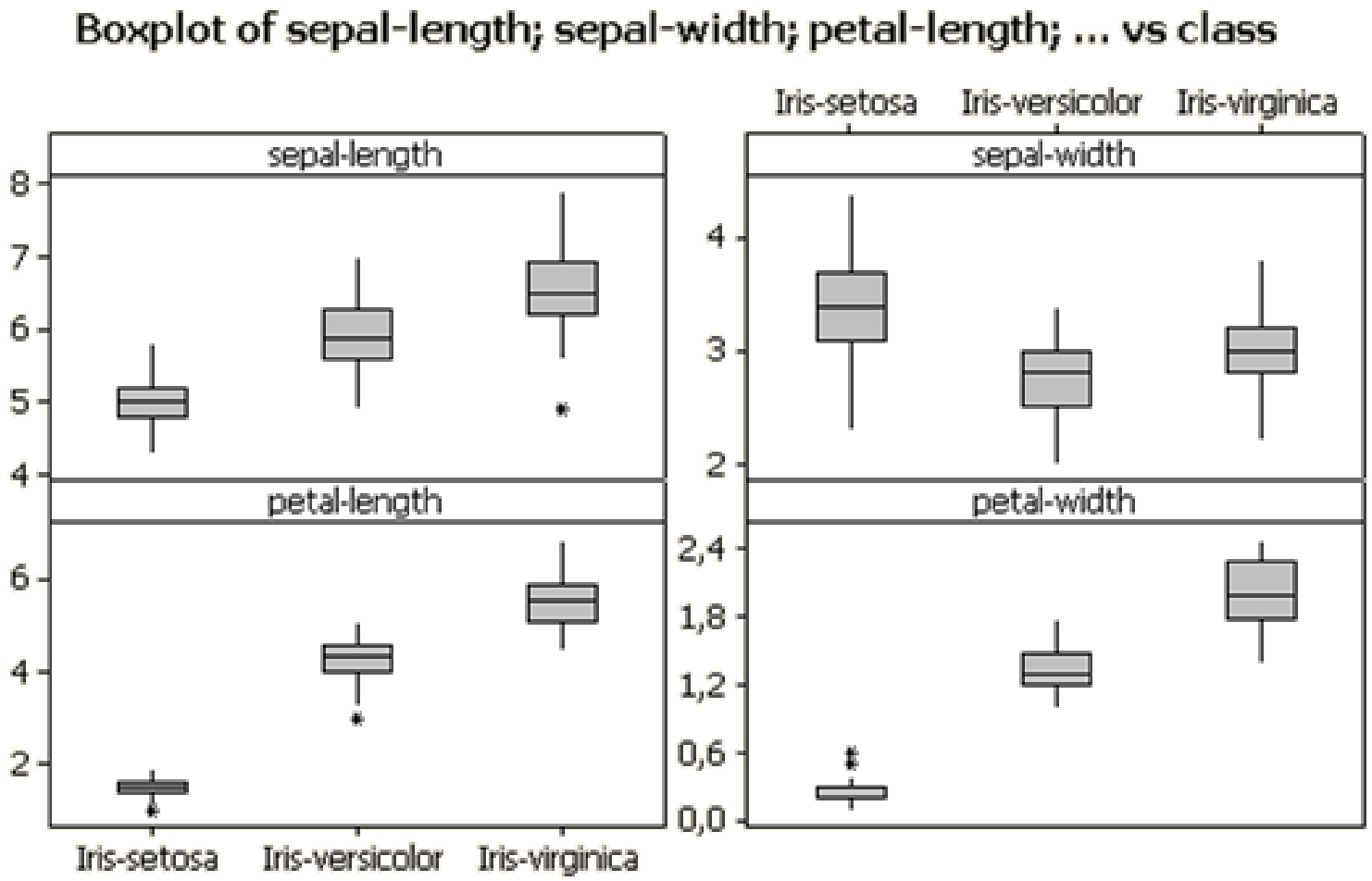
Visualization

- ✓ Box Plots



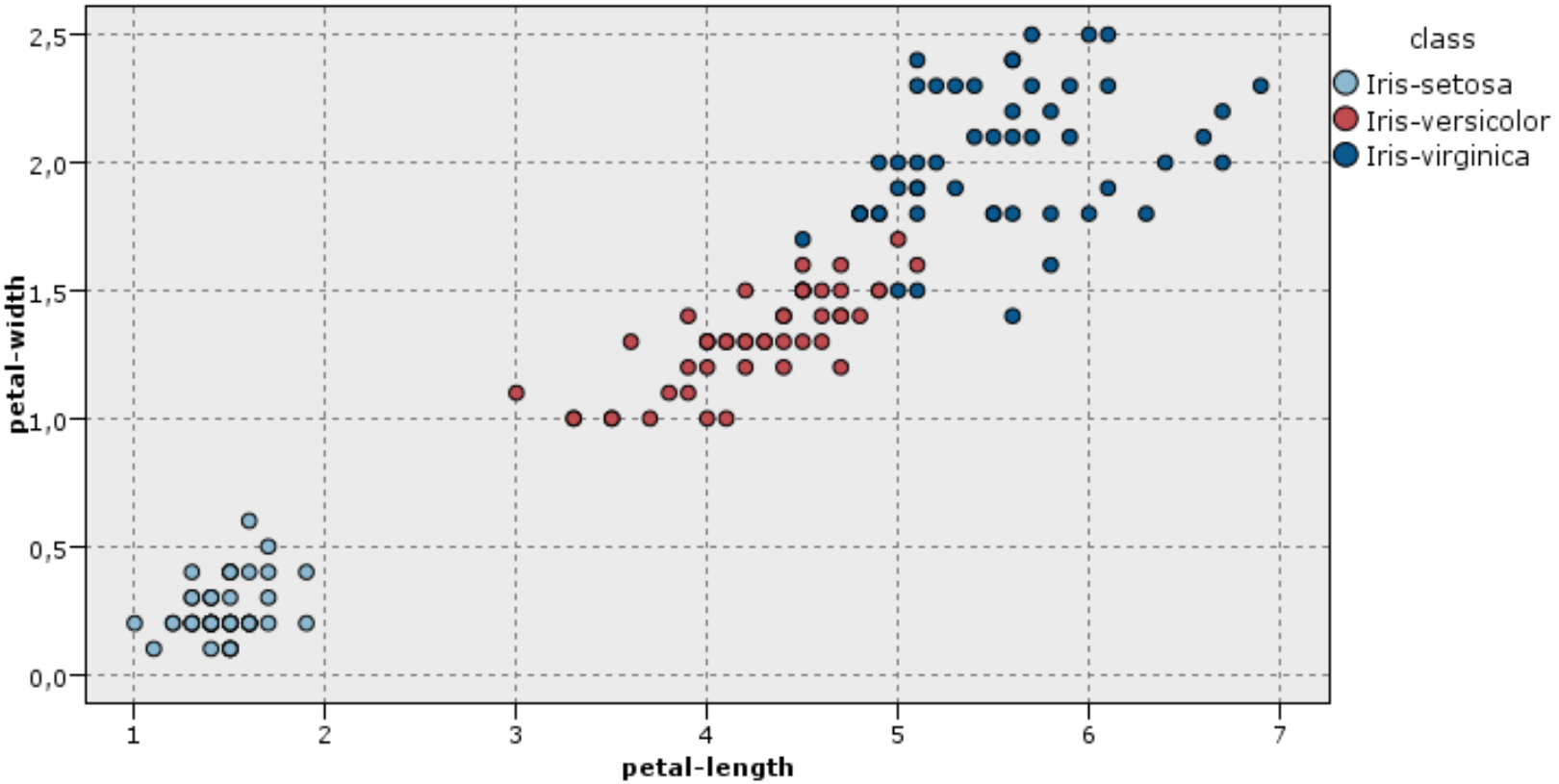
Visualization

- ✓ Box Plots



Visualization

- ✓ Scatter Plot

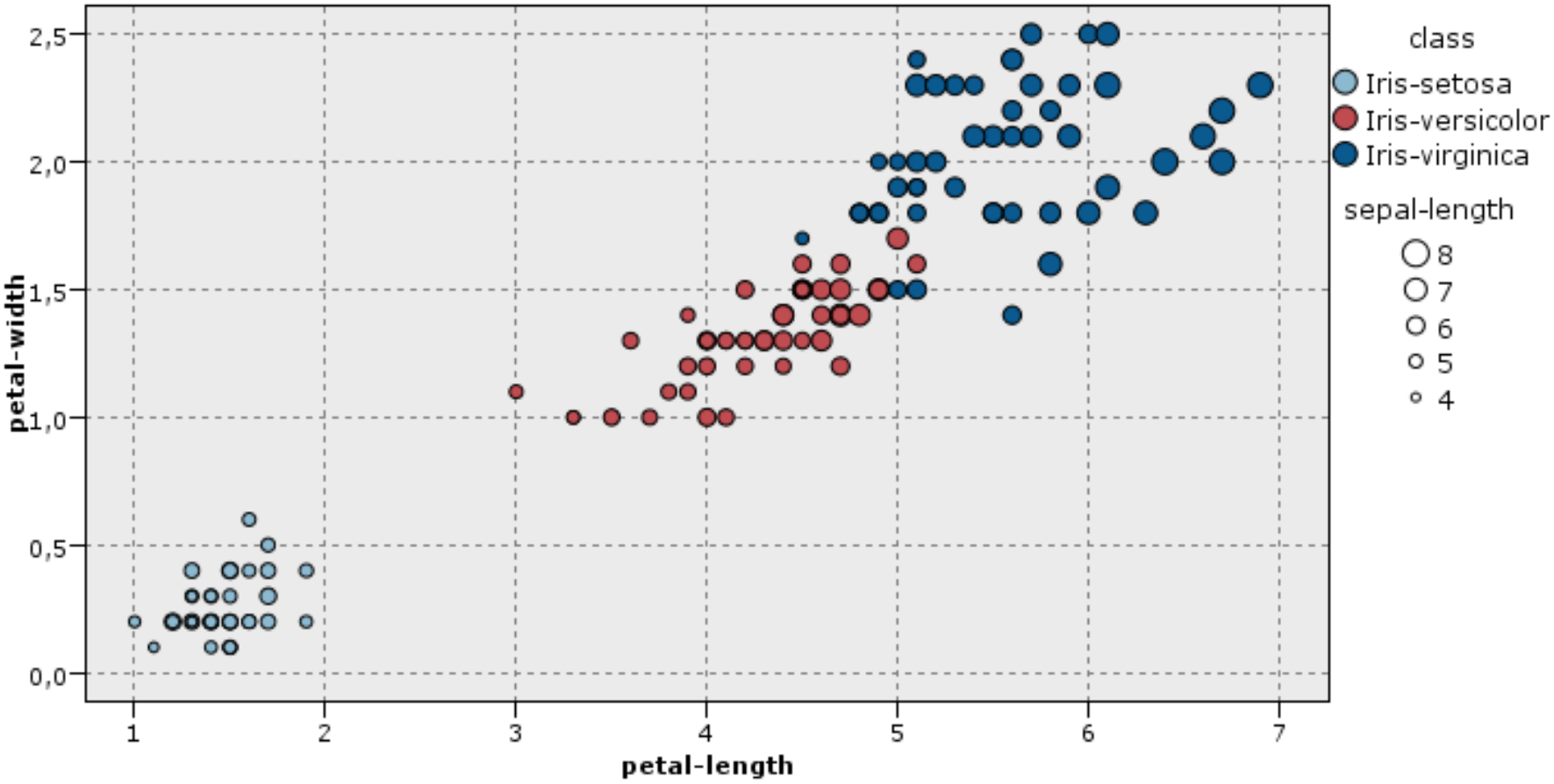


Visualization

- ✓ Scatter Plot
  - 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.

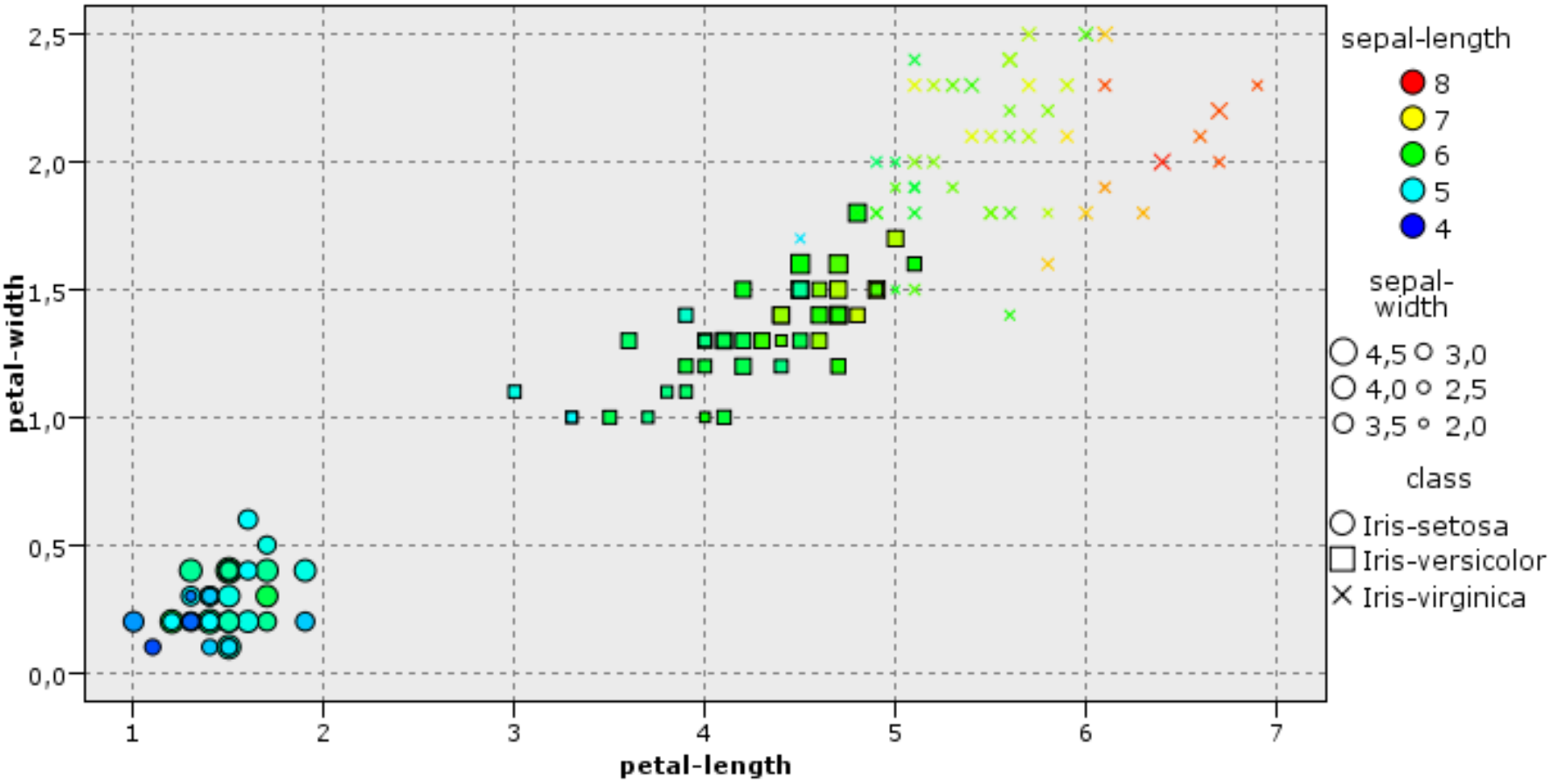
Visualization

- ✓ Scatter Plot



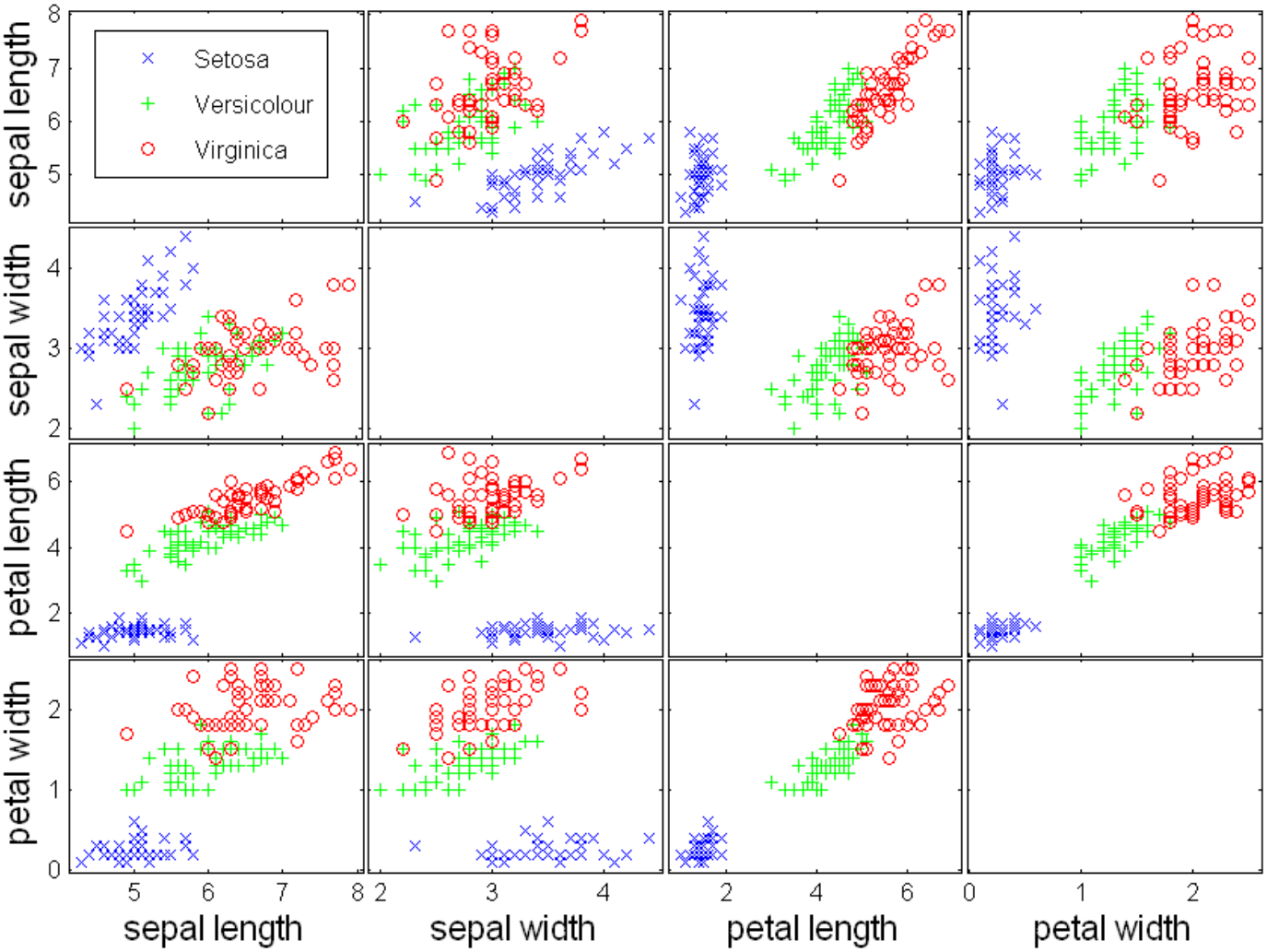
Visualization

- ✓ Scatter Plot



Visualization

- ✓ Array of Scatter Plot

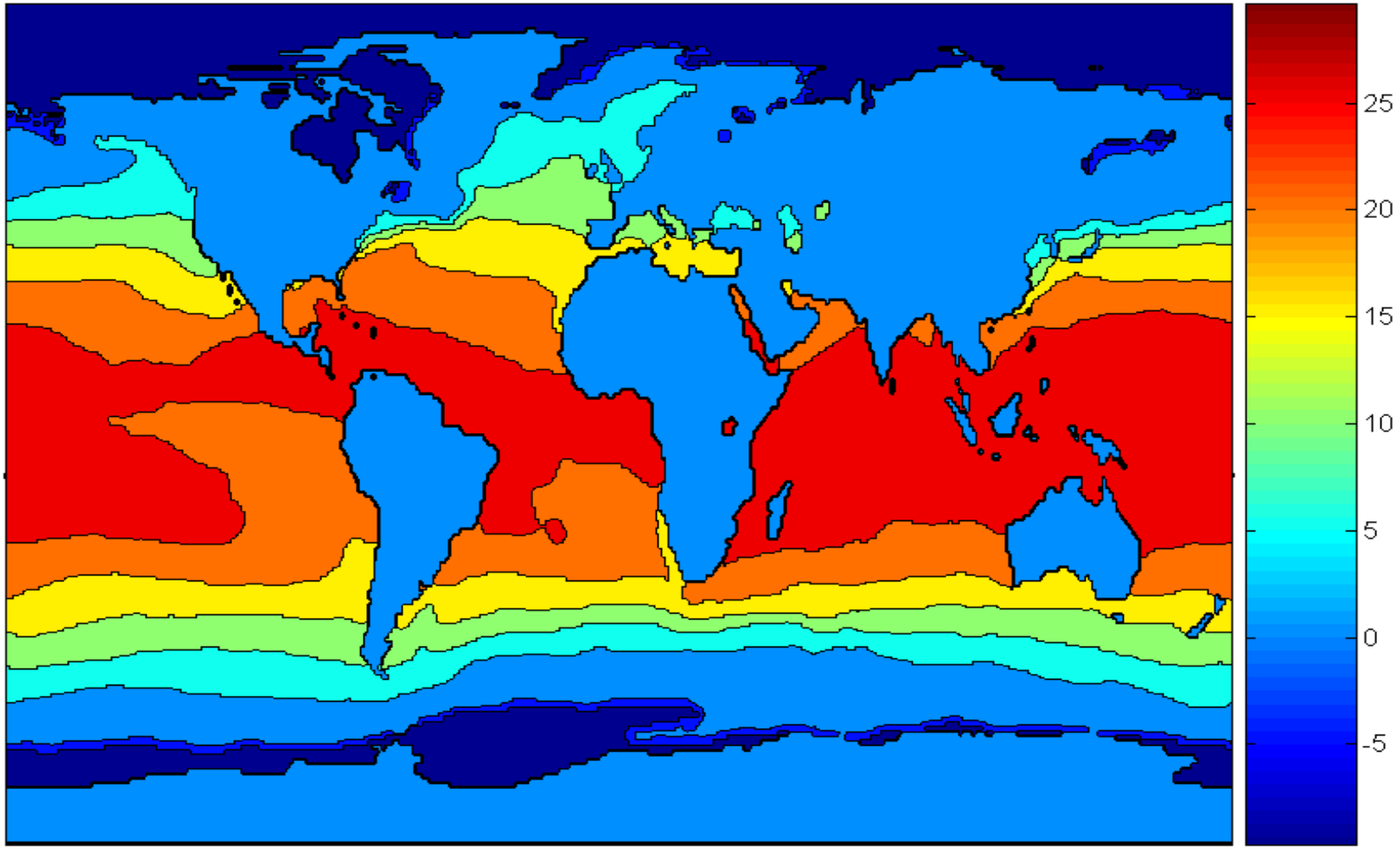


Visualization

- ✓ 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.

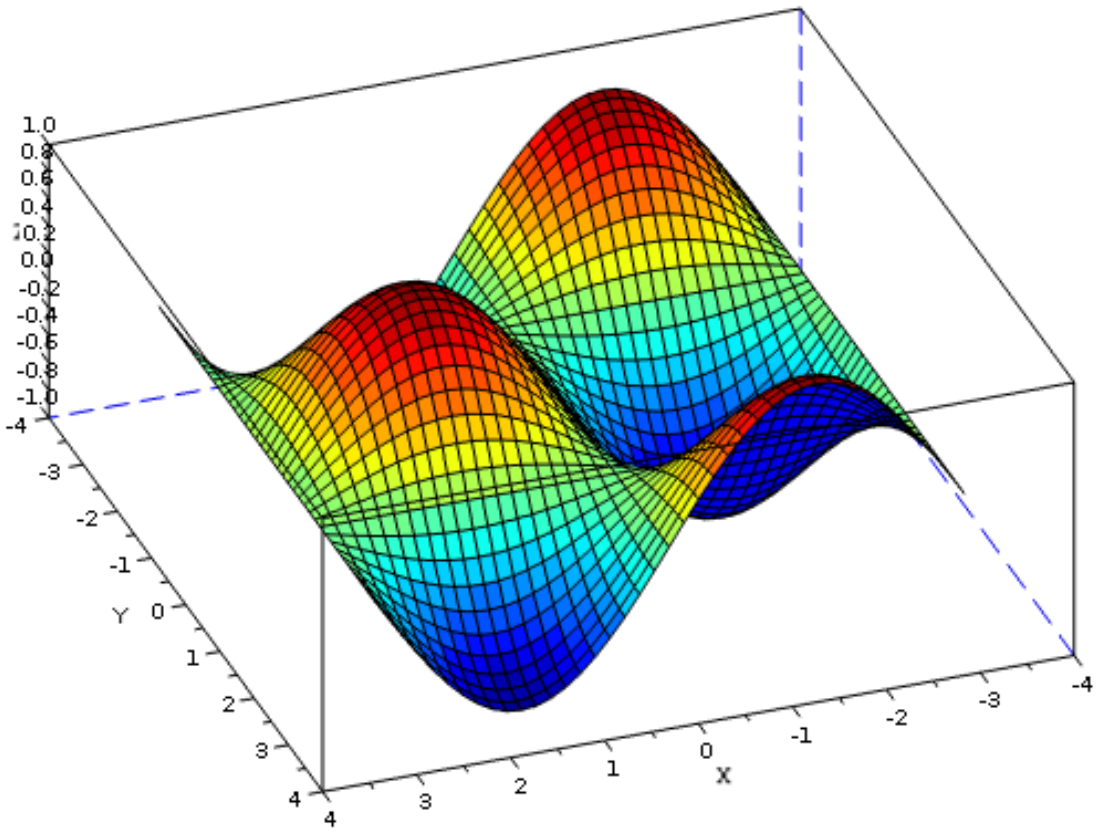
Visualization

- ✓ Contour plots example: Sea Surface Temperature – December 1998



Visualization

- ✓ Surface Plot

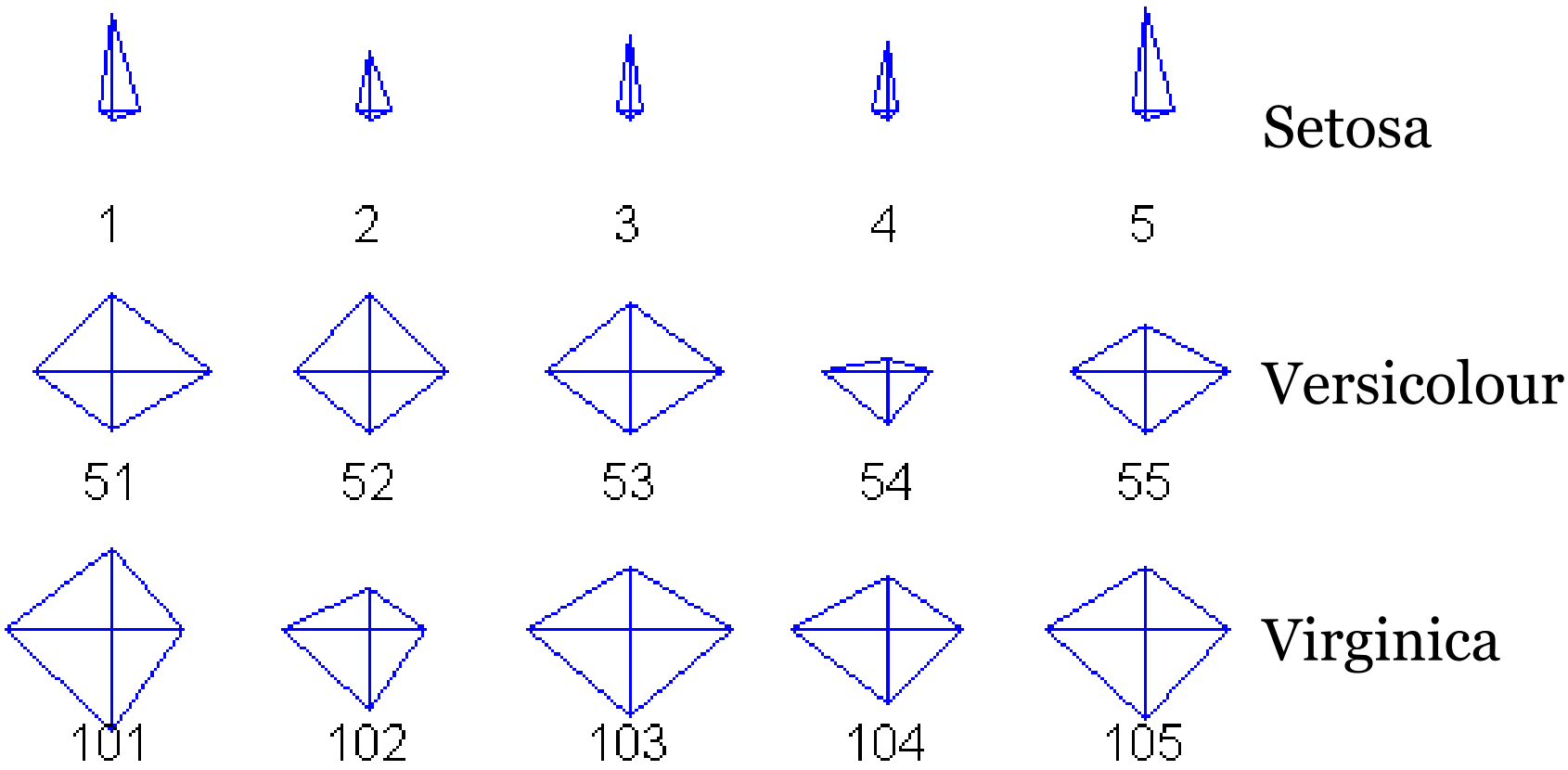


Visualization

- ✓ Star Plots
  - This technique use one axis for each attribute,
  - The 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

Visualization

- ✓ Star Graph for 15 Iris flowers



Visualization

- ✓ Chernoff Faces for 15 Iris flowers