Outline of Basic data exploration techniques

- Data properties
  - Attributes and Objects
  - Types of Data
  - Data Quality
- Basic data exploration techniques
  - Basic statistics
  - Data visualization

What is data exploration?

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
  - Seminal book is Exploratory Data Analysis by Tukey

John Tukey

John Wilder Tukey

Born	June 16, 1915 New Bedford, Massachusetts, U.S.
Died	July 26, 2000 (aged 85) New Brunswick, New Jersey, U.S.
Nationality	American
Alma mater	Brown University Princeton University
Known for	Exploratory data analysis Projection pursuit Box plot Cooley–Tukey FFT algorithm

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory

- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization

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

  - Most summary statistics can be calculated in a single pass through the data

Frequency and 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 an attribute is the most frequent attribute value (e.g. 1,2,3,2,24,3,2: mode=2)
- The notions of frequency and mode are typically used with **categorical data**

Percentiles

- For **continuous data**, the notion of a percentile is more useful.

https://www.mathsisfun.com/data/percentiles.html

- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.
- Percentiles are commonly used to report scores in tests, like the SAT, GRE and LSAT. for example, the 70th percentile on the 2013 GRE was 156. That means if you scored 156 on the exam, your score was better than 70 percent of test takers.

Percentiles

- 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.
  - We use the nearest-rank method to computer percentiles
  - The ordinal rank n = N: number of samples/objects. P is the percentile
  - The percentile value is the nth number in the ordered list
- What is the 5th, 50th percentile of the list {15, 50, 20, 35, 40}?
  - 5th: n= = =1, so the 5th percentile is 15.
  - 50th: n= =3, so the 50th percentile is 35

In-class exercise

- Provide a set of integers so that:
  - Its mode is 5
  - Its 10th percentile is 1
  - Its 50th percentile is 5
  - Its 90th percentile is 8

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
  - What is the mean of 2.2, 2.2, 2.3, 2.4, 2.5, 4.0?
- Thus, the median is also commonly used.

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

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

$X_{(r+1)}$: the (r+1)th number in the ranked array X

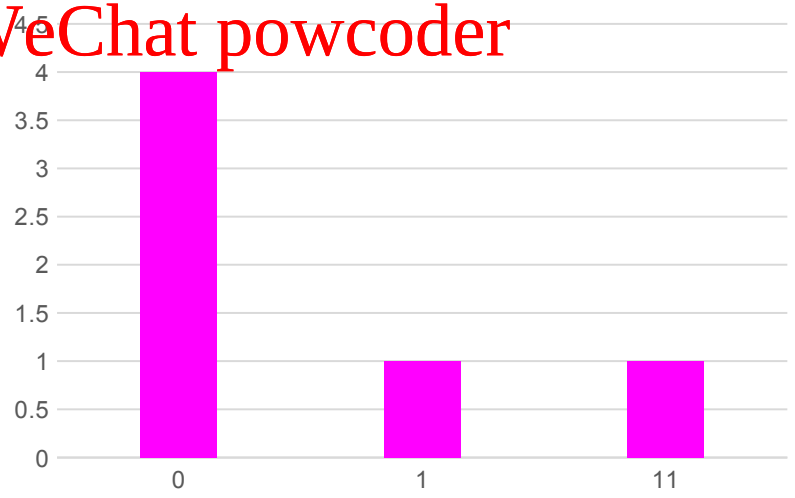
X=(1,2,3,9,15) ➔ X_1=1, X_2=2, X_4=9

Review of mean, median, and mode

- **Mean:** The "average" number; found by adding all data points and dividing by the number of data points.
  - Example: The mean of 4, 11, and 7 is (4+1+7)/3.
- **Median:** The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).
  - Example: The median of 4, 1, and 7 is 4
- **Mode:** The most frequent number—that is, the number that occurs the highest number of times.
  - Example: The mode of 4, 4, 3, 2, 2, 2 is 2 because it occurs three times, which is more than any other number.

In-class exercises

- Give an array of integer A such as A's mean = A's median = A's mode
- Give an array of integer A such as A's mean < A's median
- Give an array of integer A such as A's mean > A's median
  - 0, 0, 0, 0, 1, 11



Measures of Spread: Range and Variance

- Range is the difference between the max and min

- The variance or standard deviation s_x is the most common measure of the spread of a set of points.

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

Outline of Lecture 2

- Data properties
  - Attributes and Objects
  - Types of Data
  - Data Quality
- Basic data exploration techniques
  - Basic statistics
  - Data visualization
- Introduction to classification problems
  - Decision tree

Visualization

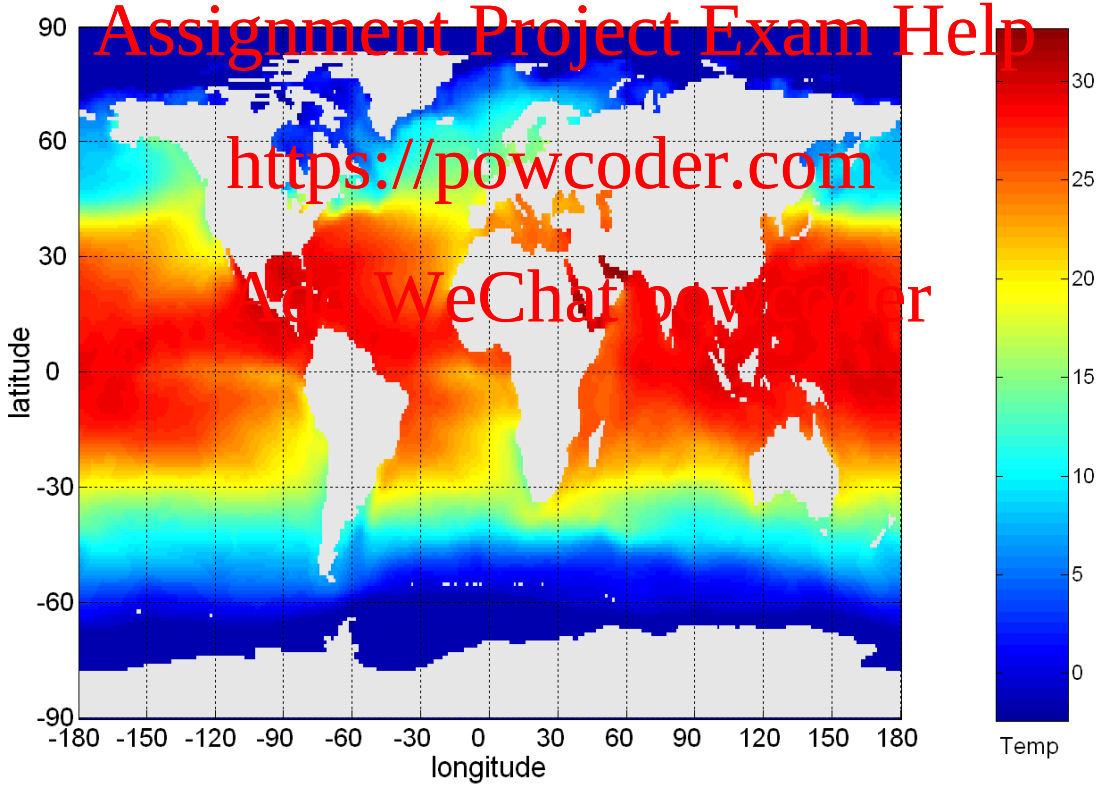
Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- 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

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) for July 1982
  - Thousands of data points are summarized in a single figure



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.
- 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.

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

- Is the elimination or the de-emphasis of certain objects and attributes
- Selection may involve 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
- Selection 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

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

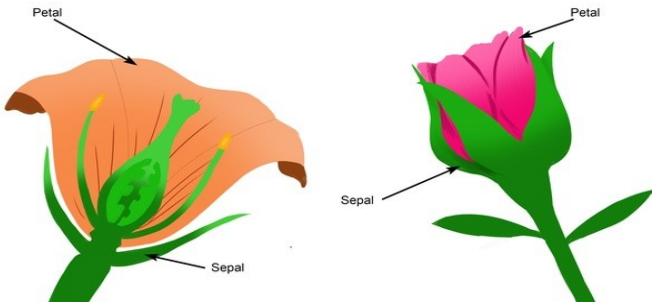
Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Iris data set

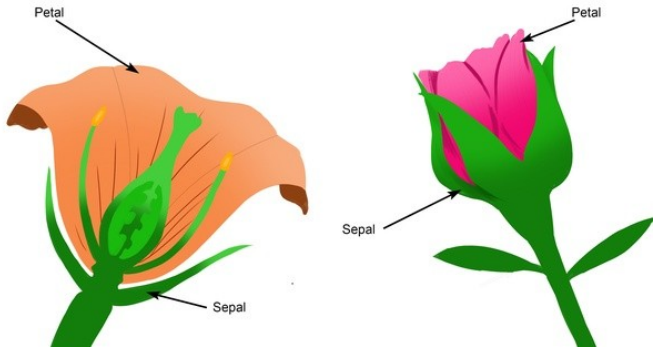
Iris Versicolor **Iris Setosa** **Iris Virginica**

from Machine Learning in R for beginners

The ***Iris* flower data set** or **Fisher's *Iris* data set** is introduced by the British statistician and biologist Ronald Fisher in his 1936 paper: ”The use of multiple measurements in taxonomic problems”.

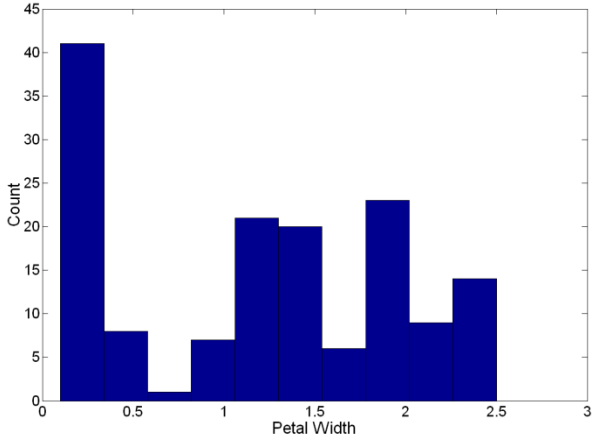


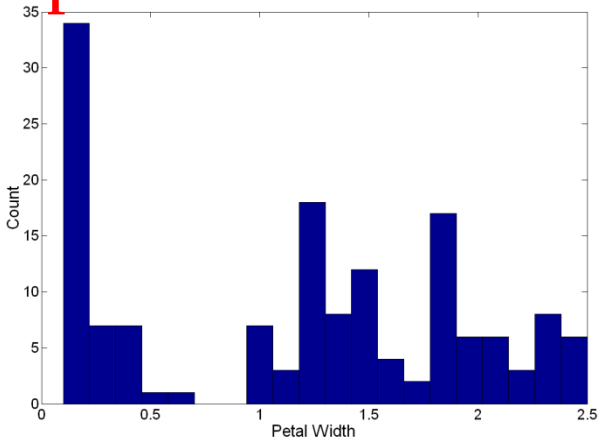
a better look at the flower



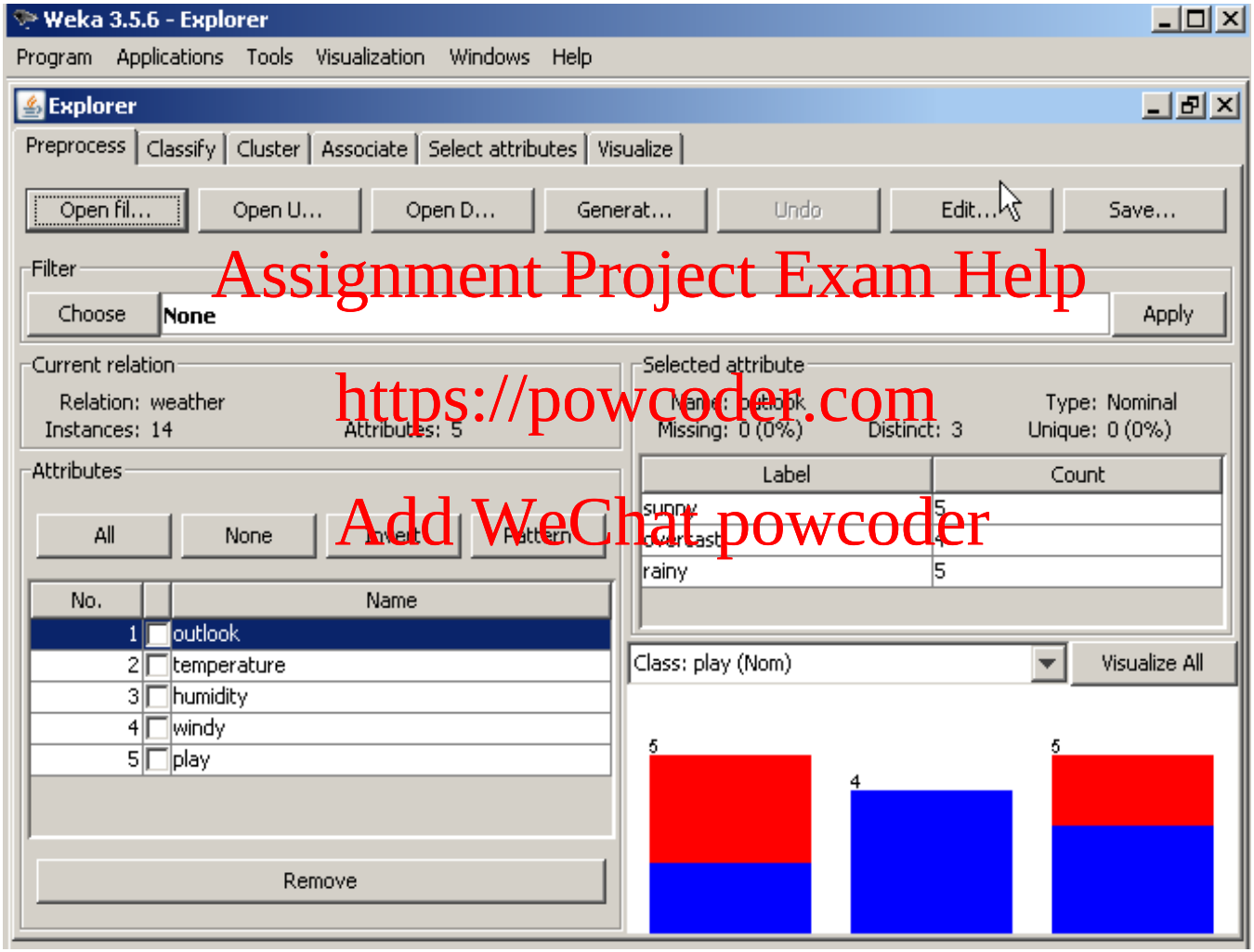
Visualization Techniques: Histograms

- 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)



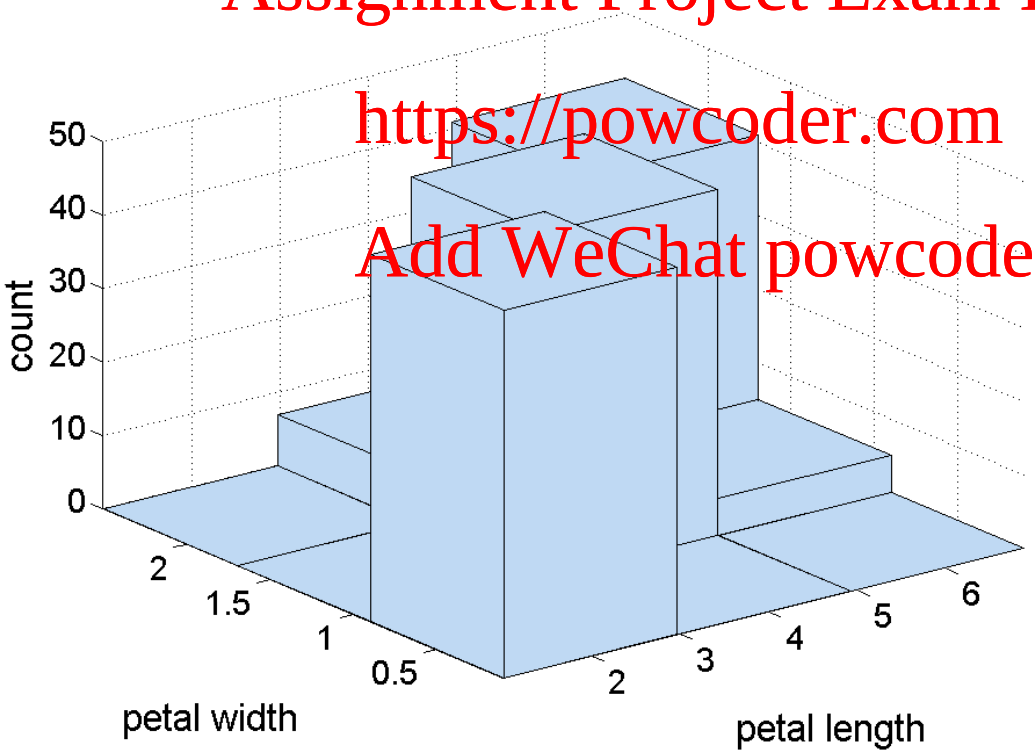


Histogram from Weka



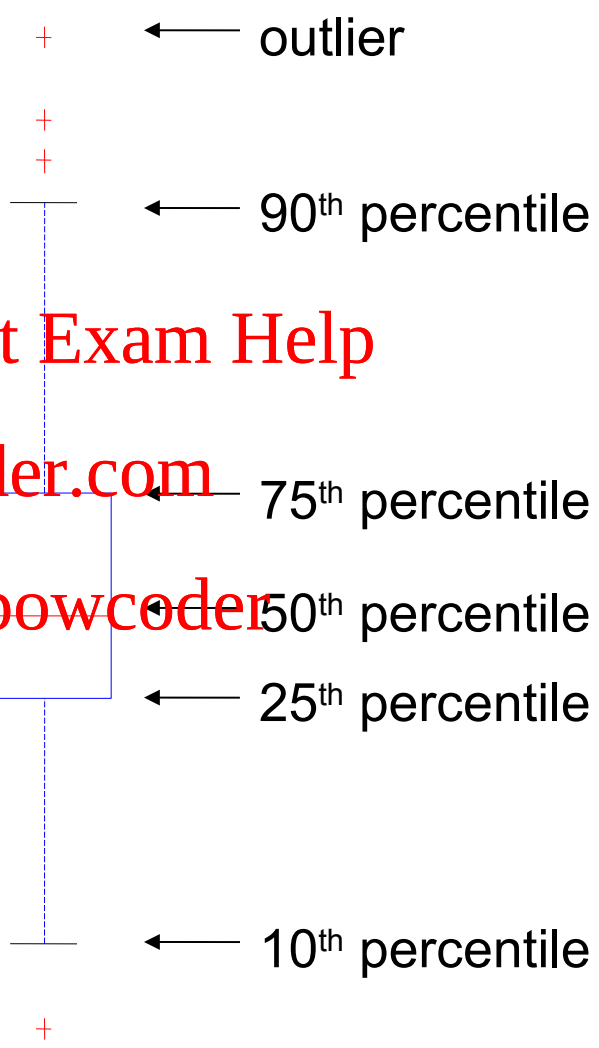
Two-Dimensional Histograms

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



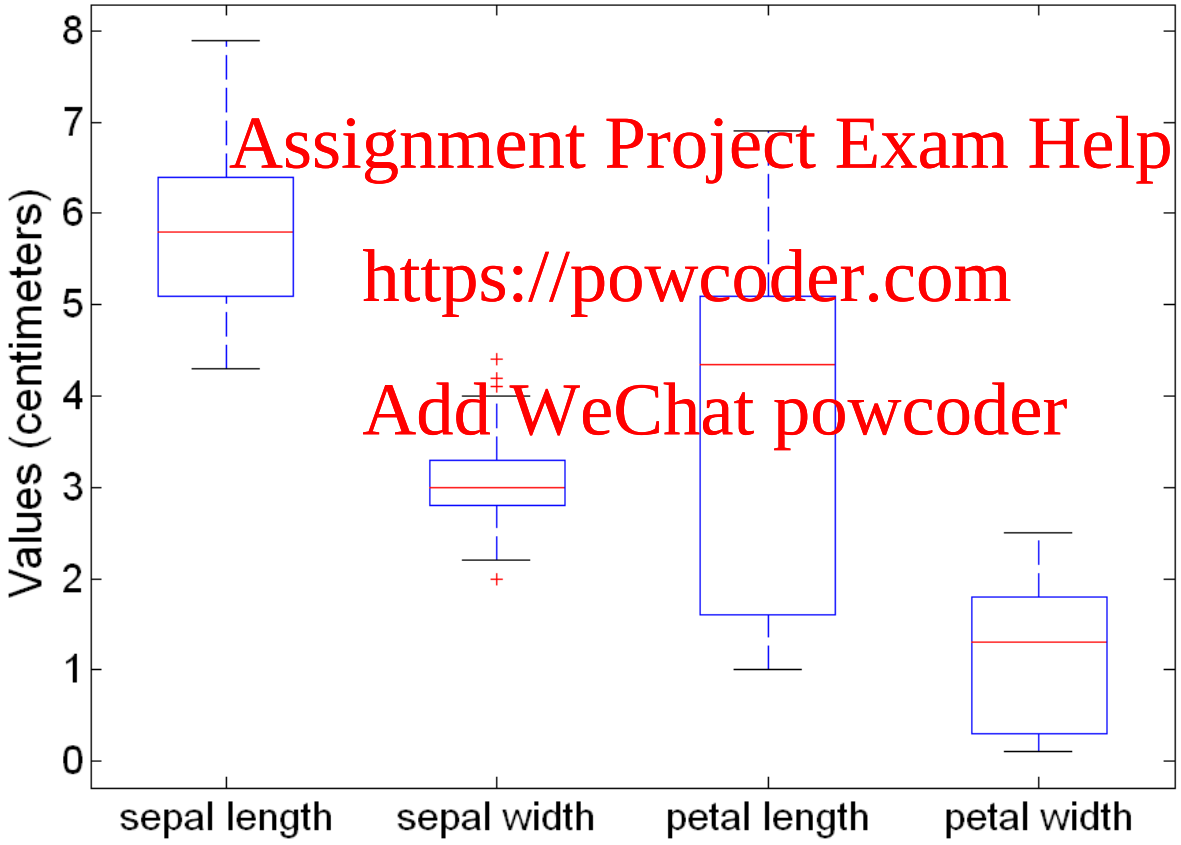
Visualization Techniques: Box Plots

- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



Example of Box Plots

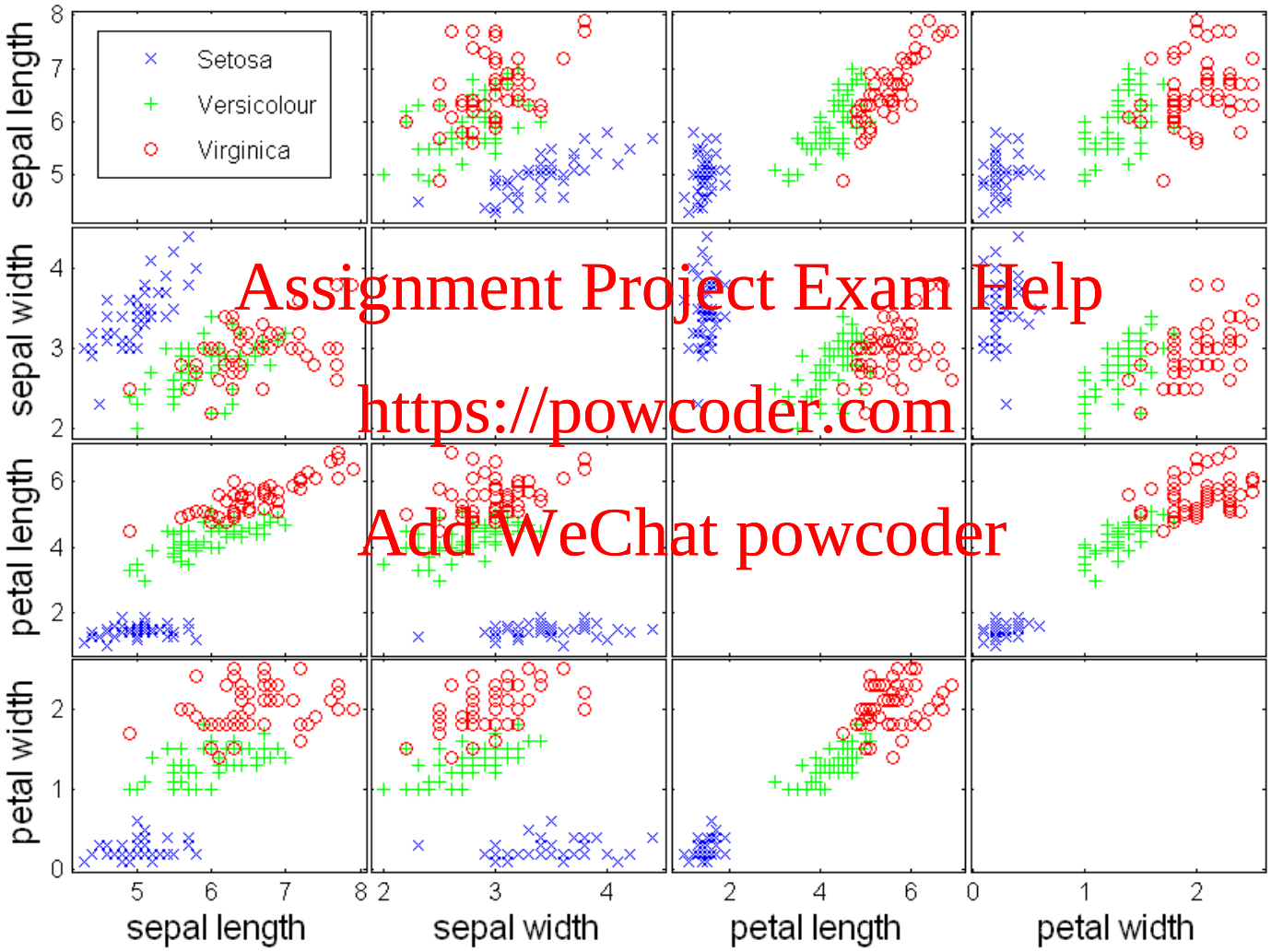
- Box plots can be used to compare attributes



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
    - See example on the next slide

Scatter Plot Array of Iris Attributes



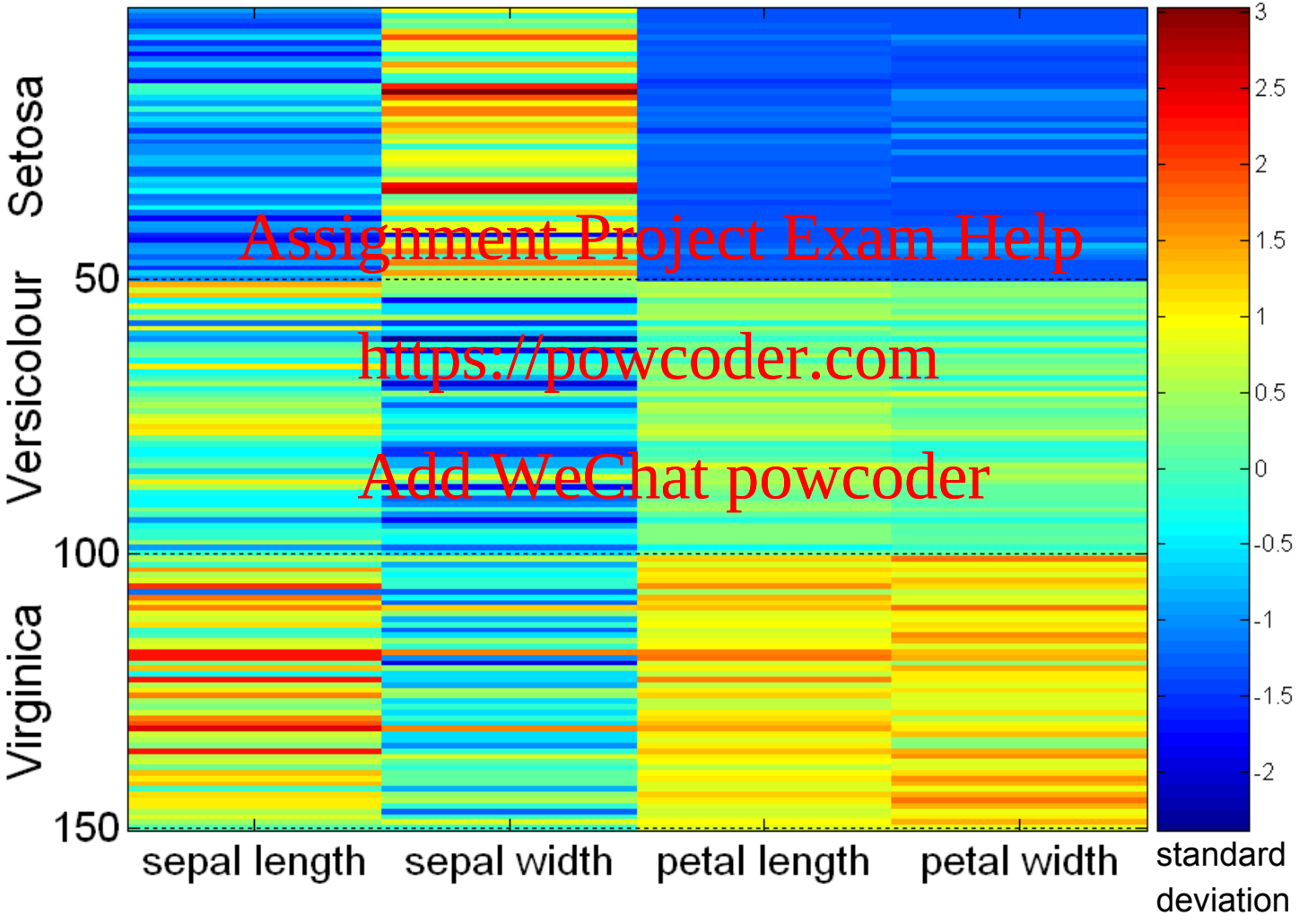
Visualization Techniques: Matrix Plots

- 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

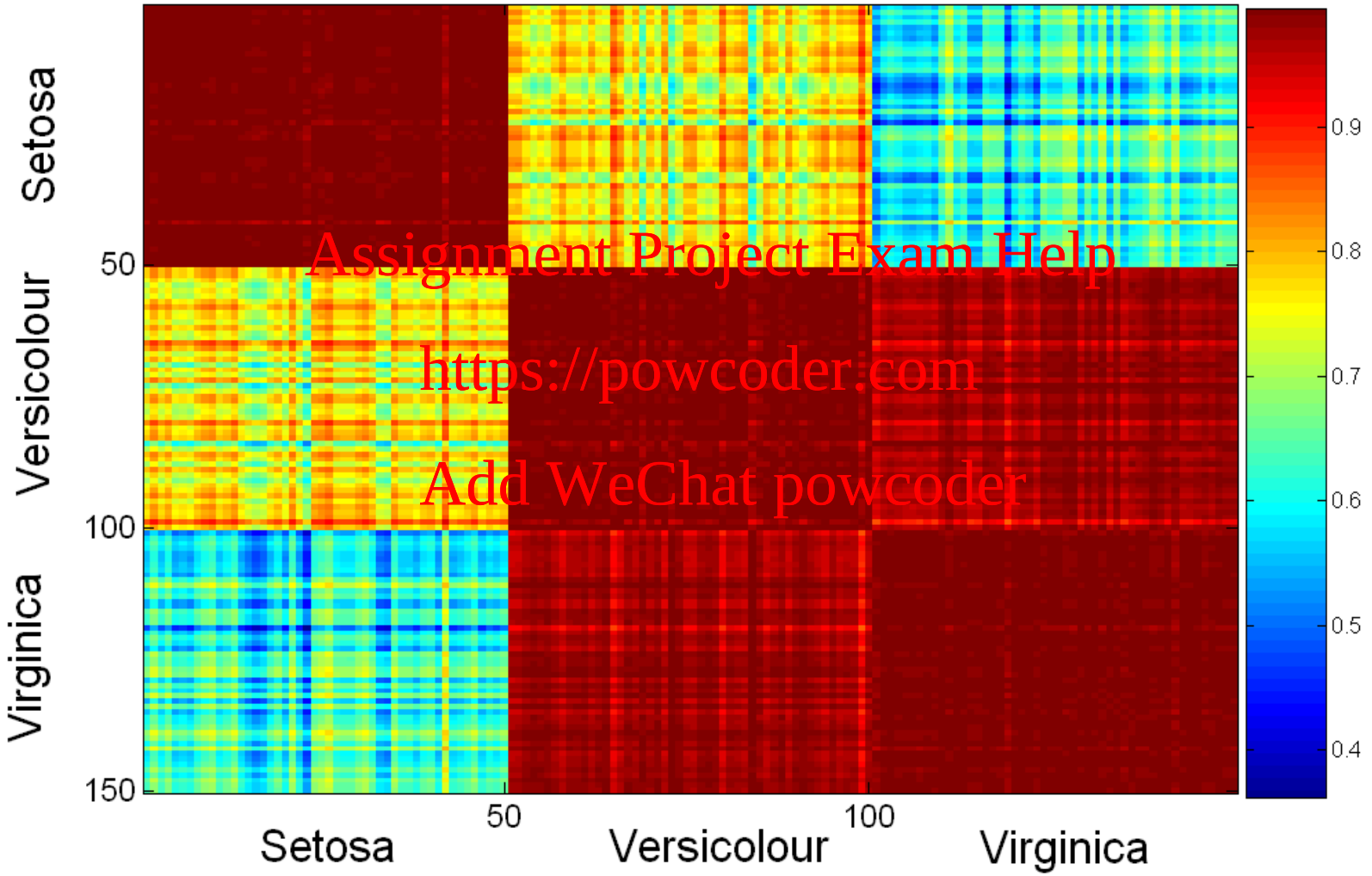
Matrix plot -continued

- An image can be regarded as a rectangular array of pixels
- Each pixel is characterized by its color and brightness
- A data matrix can be visualized as an image by associating each entry of the data matrix with a pixel in the image
  - The brightness or color of the pixel is determined by the value of the corresponding entry of the matrix

Visualization of the Iris Data Matrix



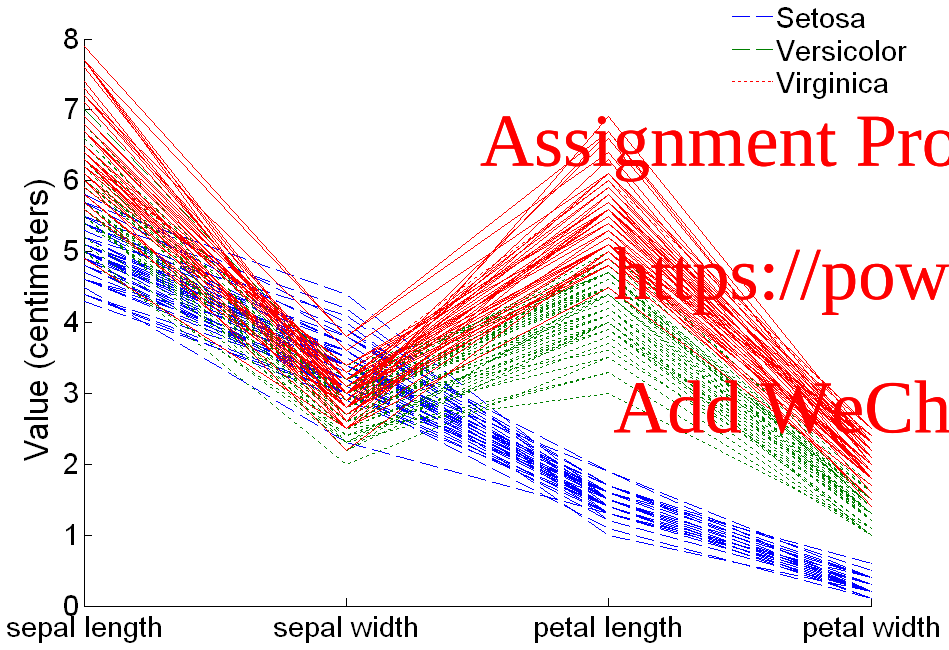
Visualization of the Iris Correlation Matrix

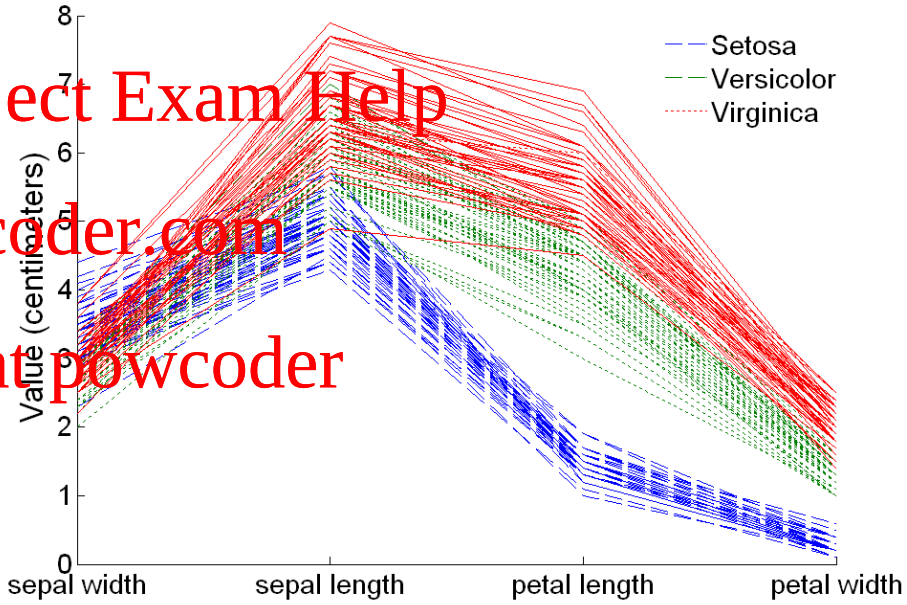


Visualization Techniques: Parallel Coordinates

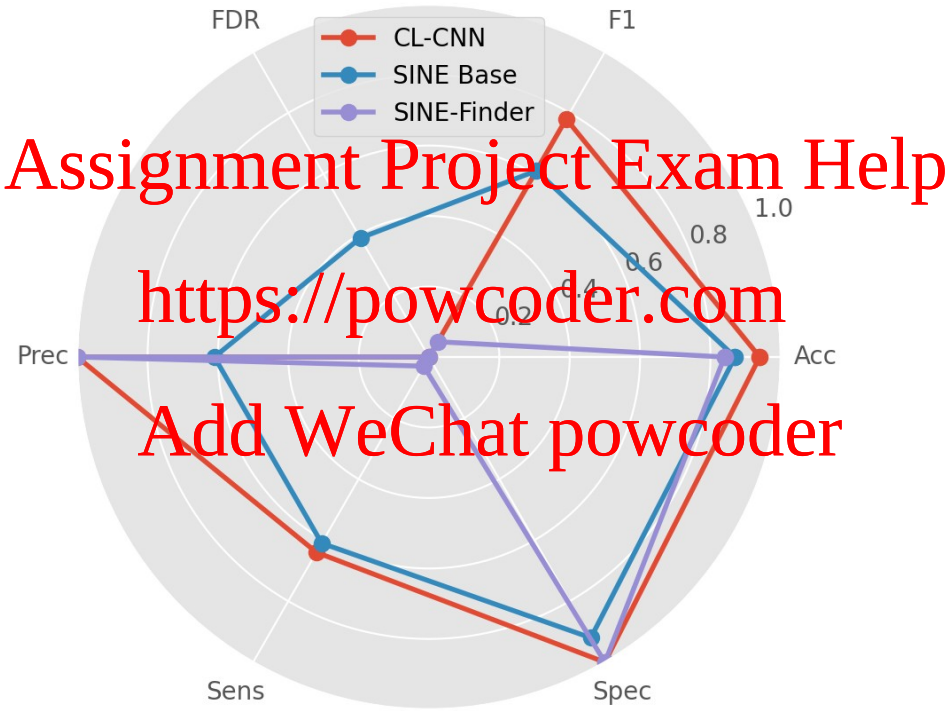
- Parallel Coordinates
  - Used to plot the attribute values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus,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





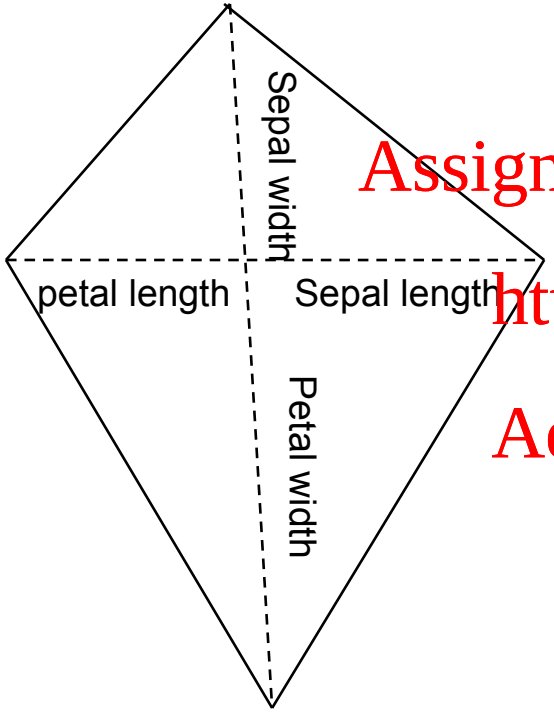
Radar chart



Other Visualization Techniques

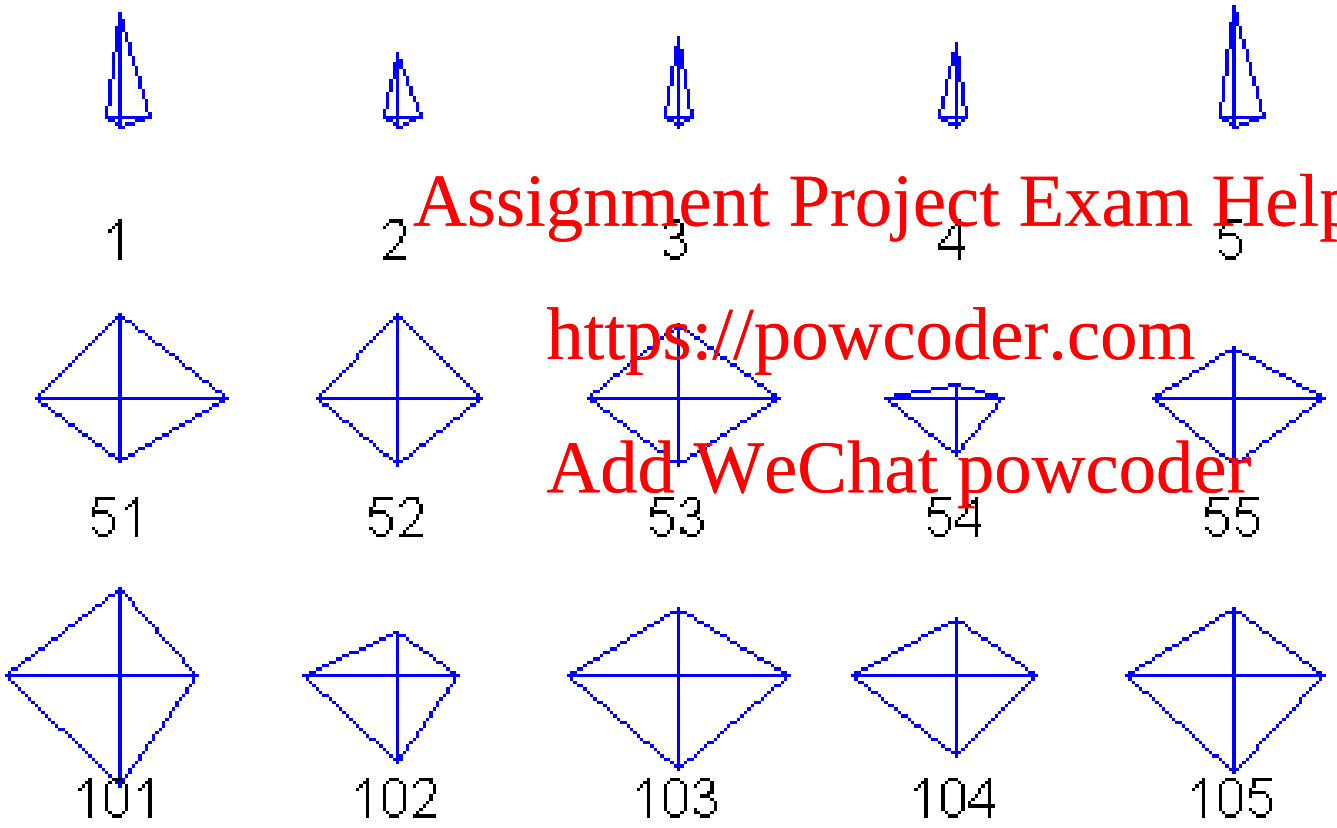
- Star Plots
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon
- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values 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 coordinate graph



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

Star Plots for Iris Data



Setosa

Versicolour

Virginica

Creating a Multidimensional Array

- Converting tabular data into a multidimensional array:
  - Identify which attributes are to be the dimensions and which attribute is to be the target attribute
    - Values of target variable appear as entries in the array
    - The target value is typically a count or continuous value
    - Can have no target variable at all except the count of objects that have the same set of attribute values
  - Find the value of each entry in the multidimensional array by summing the values (of the target attribute) or the count of all objects that have the attribute values corresponding to that entry.

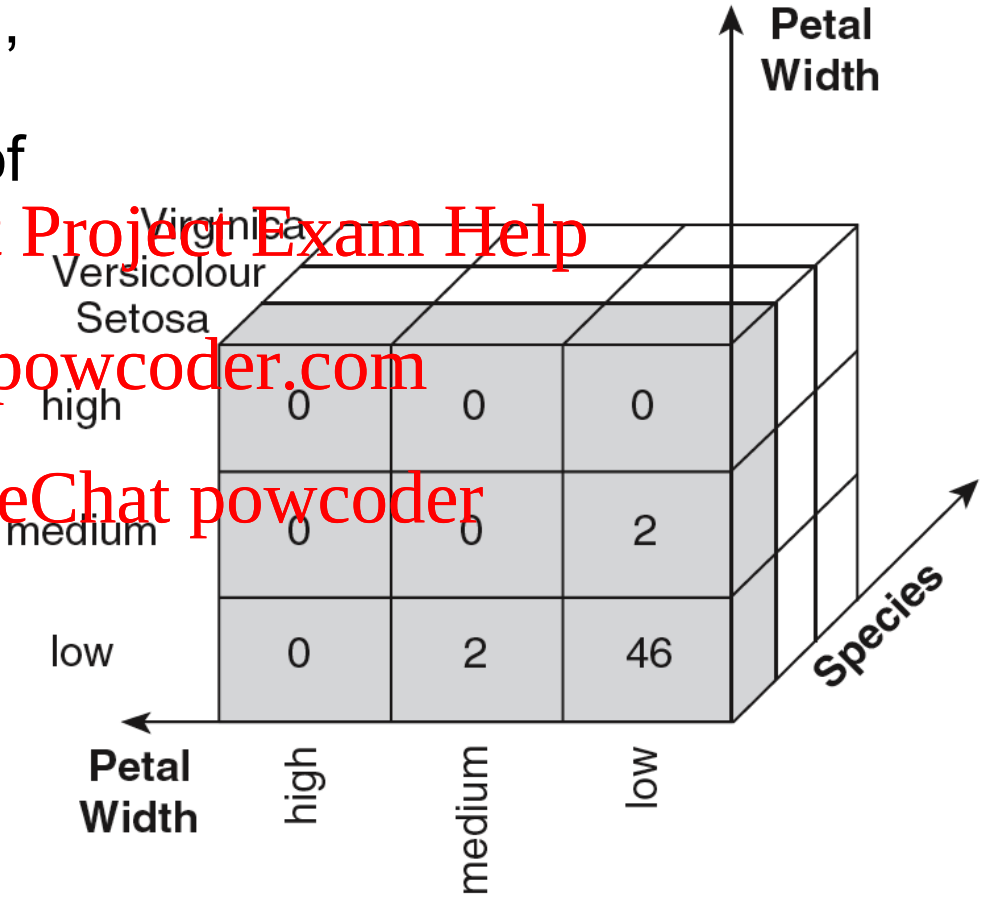
Example: Iris data

- We show how the attributes, petal length, petal width, and species type can be converted to a multidimensional array
  - First, we discretized the petal width and length to have categorical values: *low*, *medium*, and *high*

Petal Length	Petal Width	Species Type	Coun
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Example: Iris data (continued)

- Each unique tuple of petal width, petal length, and species type identifies one element of the array.
- This element is assigned the corresponding count value.
- The figure illustrates the result.
- All non-specified tuples are 0.



Example: Iris data (continued)

- Slices of the multidimensional array are shown by the following cross-tabulations
- What do these tables tell us?

Length	Width: low	Width: medium	Width: high
low	46	2	0
medium	2	0	0
high	0	0	0

Length	Width: low	Width: medium	Width: high
low	0	0	0
medium	0	43	3
high	0	2	2

Length	Width: low	Width: medium	Width: high
low	0	0	0
medium	0	0	3
high	0	3	44