



Data Warehouse and Data Mining

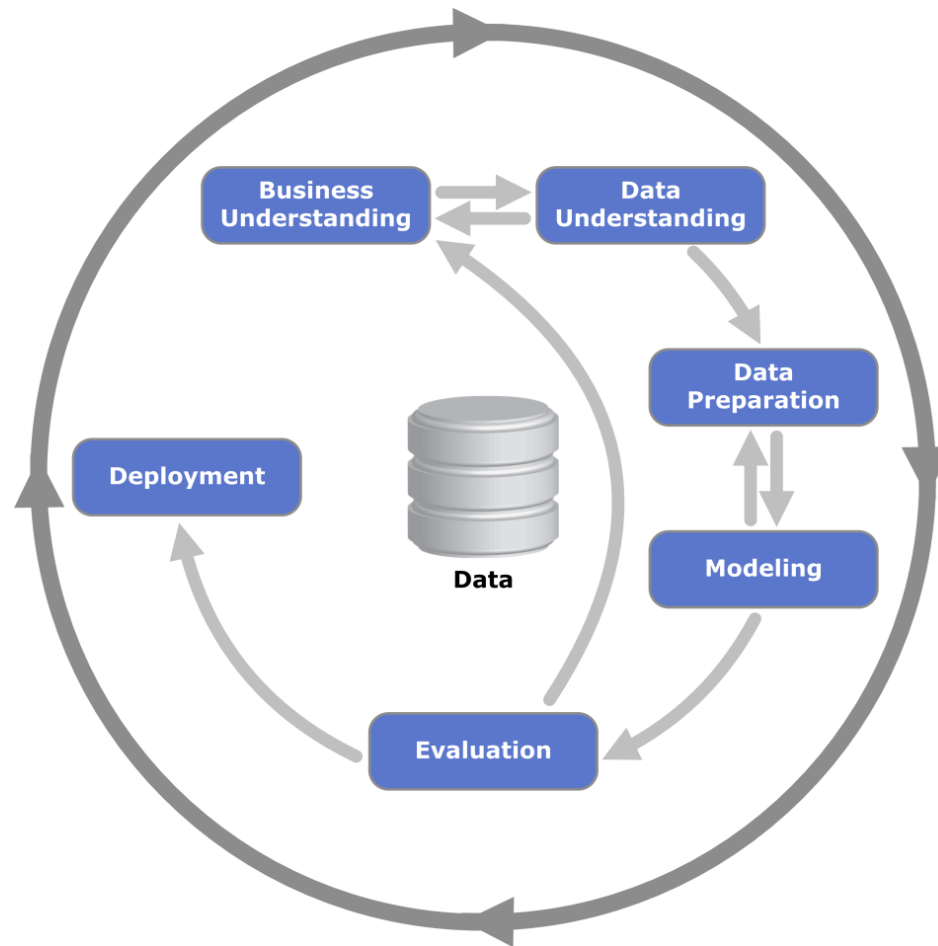
Module II – Data Mining

Data Understanding

Ph.D. Ettore Ritacco



The Knowledge Discovery Process (CRISP-DM)





About the Lecture

- Main Source:
 - Tan, Steinbach, Kumar “**Introduction to Data Mining**”



Data - Objects and Attributes

- *Data* is a collection of objects
- *Objects* (a.k.a. elements, instances, samples, records, rows, ...) are described by means of a set of attributes
- An *attribute* (a.k.a. field, variable, feature, ...) defines a property, a characteristic or a measure of an object (e.g. eye color, temperature...)



Data – An example

ID	Age	Sex	Marital Status	Income	Job	Trustable
23432	34	male	single	24000	bank officer	yes
23433	45	male	-	36000	Teacher	yes
23434	44	female	single	300000	-	no
23435	55	female	divorced	35000	Unemployed	no
23436	57	female	married	22000	bank officer	no



Attribute Types

- Categorical (Qualitative) Attributes [Discrete attributes]
 - Nominal
 - Ordinal
- Numeric (Quantitative) Attributes [Discrete and Continuous attributes]
 - Interval-scaled
 - Ratio-scaled
- More complex data
 - Texts in natural language, Dates, Taxonomy, Graphs, ...



Categorical Attributes

Nominal: categories, states, or “names of things”

- Hair_color = {auburn, black, blond, brown, grey, red, white}
- marital status, occupation, ID numbers, zip codes

Binary

- Nominal attribute with only 2 values (0 and 1)
- They can be:
 - **Balanced**: both outcomes equally important (e.g., gender)
 - **Unbalanced**: outcomes not equally important (e.g., medical test)

Ordinal

- Values have a meaningful order (ranking) but magnitude between successive values is not known.
- Size = {small, medium, large}, {1,2,3}, grades, army rankings



Numeric Attributes

- **Quantity** (integer or real-valued)
- **Interval-based**
 - Measured on a continuous range
 - Values have order (e.g., temperature in C° or F°)
 - No evident correlation among values
- **Ratio-Scale** (e.g., temperature in Kelvin, length, counts, monetary quantities)
 - Values are multiple of a **unit of measurement**



Discrete vs. Continuous Attributes

Discrete Attribute

(E.g., zip codes, profession, ID numbers, the set of words in a collection of documents)

- Has only a **finite** or **countably infinite** set of values
- Sometime represented as **integer variables**
- Note: Binary attributes are a special case of discrete attributes

Continuous Attribute

(E.g., temperature, height, or weight)

- Has **real numbers** as attribute values
- Practically, real values can only be measured and represented using a finite number of digits
- Typically represented as **floating-point variables**



Properties of Attributes

Type	Properties	Transformations	Operations
Nominal	Distinctness (= & ≠)	Permutations	Mode, entropy, correlation...
Ordinal	Order (< & >)	Order preserving change of values	Median, precentiles....
Interval	Addition (+ & -)	$new_value = a + old_value * b$	Mean, St. Dev, Pearson's correlat.
Ratio	Multiplication (* & /)	$new_value = a * old_value$	Geom. Mean, Harmonic Mean...

- Each type possesses all the properties and operations of the attribute types above it



Types of Data

- The most generic type is the **Record Data**
- Other types:
 - **Text Data** (corpora of documents written in natural language)
 - **Graph Data** (used to represent information from World Wide Web or Molecular Structures)
 - **Ordered Data** (e.g. Spatial Data, Temporal Data, Sequential Data, Genetic Sequence Data)
 - ...



Record Data

- It consists of a collection of records (tuples)
- Each record consists of a fixed set of attributes
- There is no explicit relationship among attributes or records
- Usually stored in flat files or relational databases.



Record Data – Example

ID	Age	Sex	Marital Status	Income	Job	Trustable
23432	34	male	single	24000	bank officer	yes
23433	45	male	-	36000	Teacher	yes
23434	44	female	single	300000	-	no
23435	55	female	divorced	35000	Unemployed	no
23436	57	female	married	22000	bank officer	no



Record Data- Special Cases

Transactional Data

- Each record involves a set of items
- Typically used to represent Market Transaction Data

<i>TID</i>	<i>Items</i>
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Equival.



TID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Binary attributes, but they can also be discrete or continuous



Record Data- Special Cases

Data Matrix

- Only *numeric attributes*
- Each record can be thought as a vector in multi-dimensional space
- Can be represented by an $m \times n$ matrix
 - Rows represent the objects and columns the attributes
 - **Advantage:** All standard matrix operations can be applied.
- It can be **sparse**: only non-zero value are important



Record Data- Special Cases

Document Data

- Used to represent a set of documents, with their terms (do you remember transactional data?)
- It is a *sparse data matrix*

	team	coach	play	ball	score	game	win	lost	Timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



Exploratory Data Analysis (EDA)

- Exploratory data analysis is an approach to analyzing data sets, to summarize their main characteristics, often with visual methods.

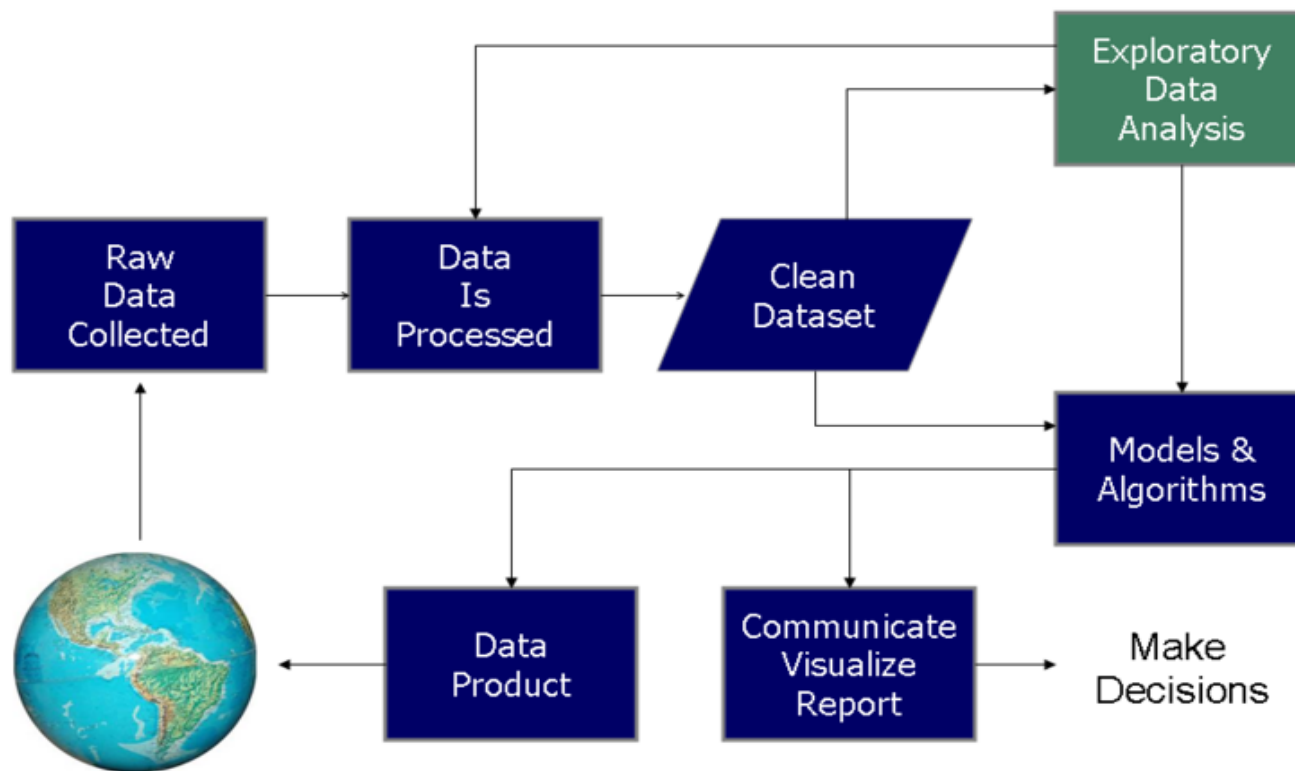
“Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”

John Tukey – “The Future of Data Analysis” - July 1961



Exploratory Data Analysis (EDA)

Data Science Process





Exploratory Data Analysis (EDA)

- Two approaches:
 - Parametric
 - The distribution, that governs the data, is known
 - Parameter estimation
 - Non-parametric
 - The distribution is unknown
 - Choose a “good” hypothesis and find its parameters



Measures for categorical attributes

- The **frequency** of an attribute value is the percentage of time the value occurs in the data set
- The **mode** of a an attribute is the most frequent attribute value
- **Variability**
 - Are there (or not) some dominant values?



Measures for numerical attributes

- Arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\begin{cases} \bar{x}_{(0)} = 0 \\ \bar{x}_{(n+1)} = \frac{n \cdot \bar{x}_{(n)} + x_{n+1}}{n+1} \end{cases}$$

- Geometric mean

$$\bar{x} = \sqrt[n]{\prod_{i=1}^n x_i}$$

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



Measures for numerical attributes

- Harmonic mean

$$\bar{x} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

- The **median** is the middle number of the group when they are ranked in order. (If there are an even number of numbers, the mean of the middle two is taken.)
 - {1, 7, 12, 23, 34, 54, 68}, the median is 23



Measures of Dispersion

- **Range:** is the difference between the max and min
- **Variance and Standard Deviation:** is the most common measure of the spread of a set of points

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

- However, this is also sensitive to outliers, so that other measures are often used.

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

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

Robust Measures for Dispersion – IQR

- Given an ordinal or continuous attribute x and a number $p \in [0,100]$, the p th *percentile* is the value of x such that $p\%$ of the observed values of x are smaller than x_p .
 - For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.
- Quartiles and outliers**
 - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
 - Inter-quartile range:
$$\text{IQR} = Q3 - Q1$$
 - Outlier:** usually, a value higher/lower than $1.5 \times \text{IQR}$

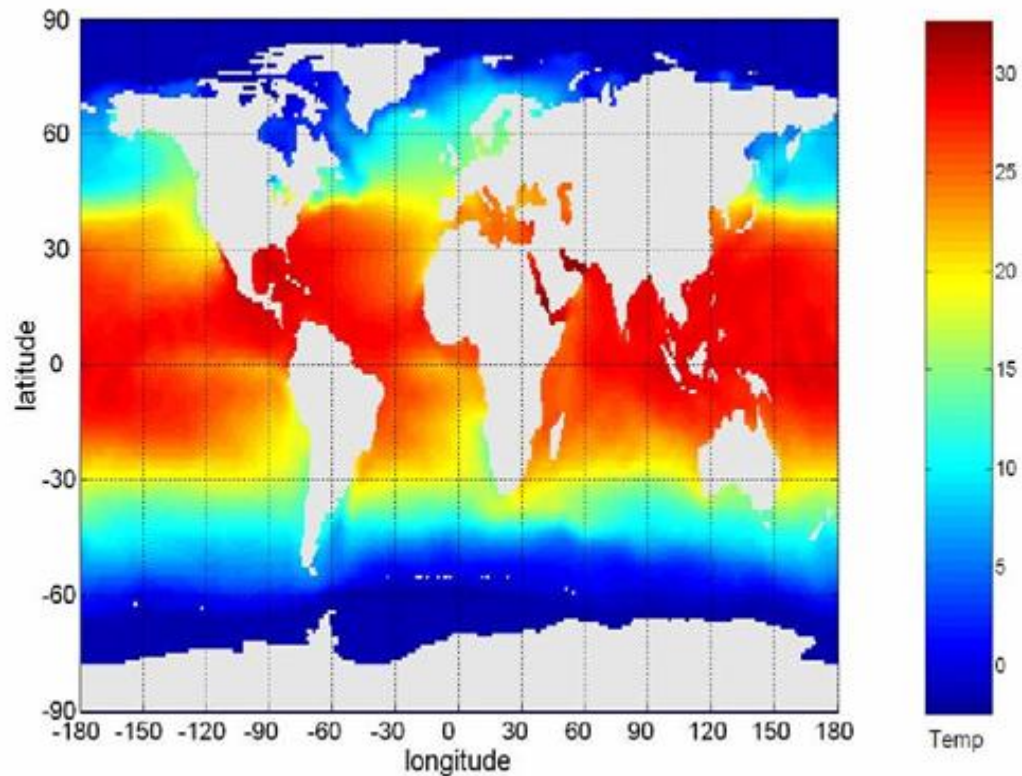


Visualization

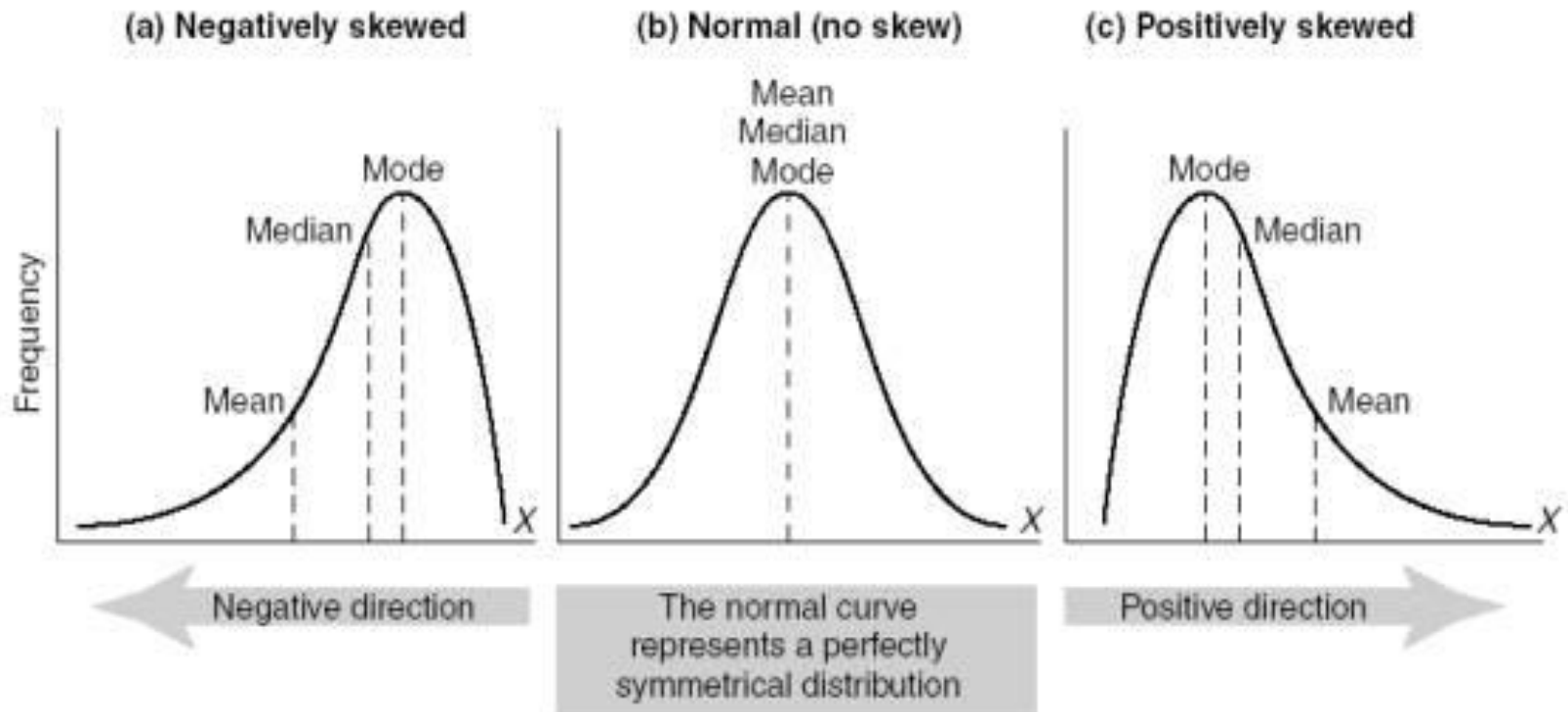
- **Aim:** To analyze/report the characteristics of the data and the relationships among data items or attributes
- **Requirement:** Conversion of data into a visual or tabular format.
- Humans have a well developed ability to analyze large amounts of information that is presented visually
 - Can detect general patterns and trends
 - Can detect outliers and unusual patterns



Visualization – Example



Normal VS Skewed Distribution





Visualization – Iris Dataset

- 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/Petal Width/Length

Setosa



Virginica

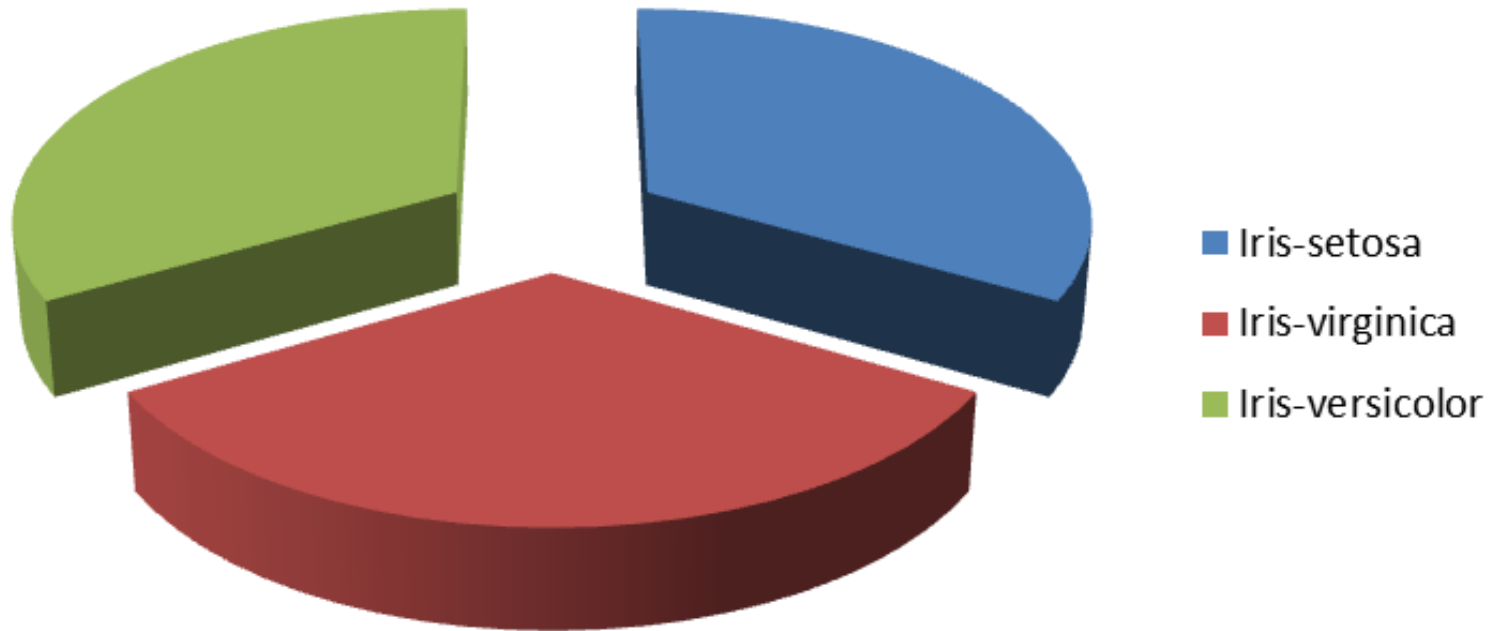


Versicolour

Visualization- Iris Dataset

sepal length	sepal width	petal length	petal width	class
4.3	3	1.1	0.1	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.4	3	1.3	0.2	Iris-setosa
4.9	2.4	3.3	1	Iris-versicolor
5	2	3.5	1	Iris-versicolor
5	2.3	3.3	1	Iris-versicolor
5.8	2.7	5.1	1.9	Iris-virginica
5.8	2.8	5.1	2.4	Iris-virginica
.....

Visualization Techniques- Pie Chart





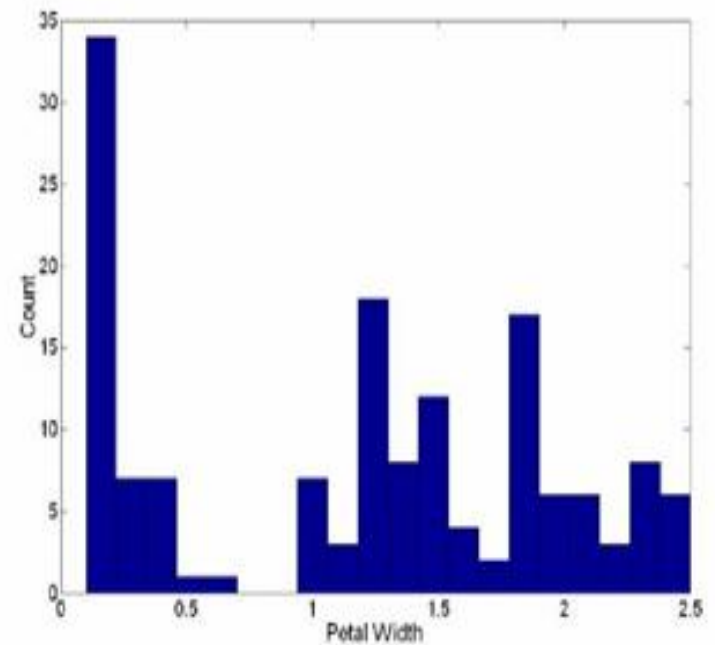
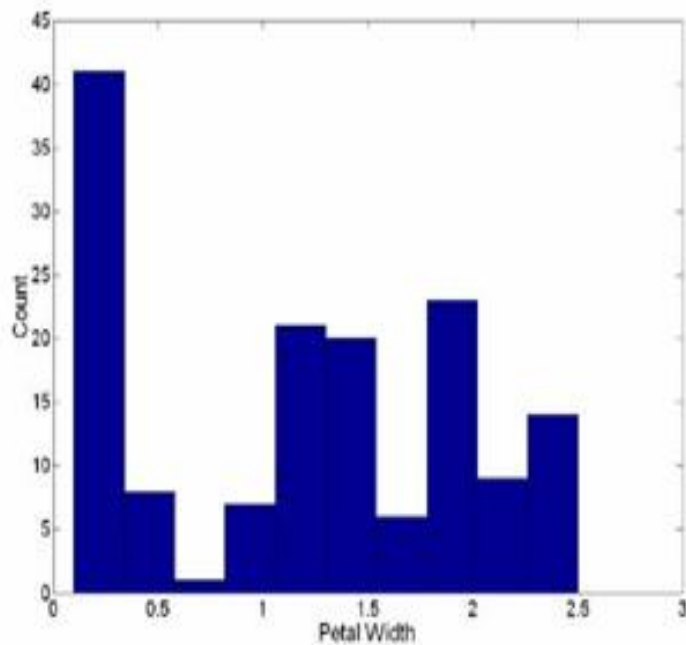
Visualization Techniques- Histogram

- 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



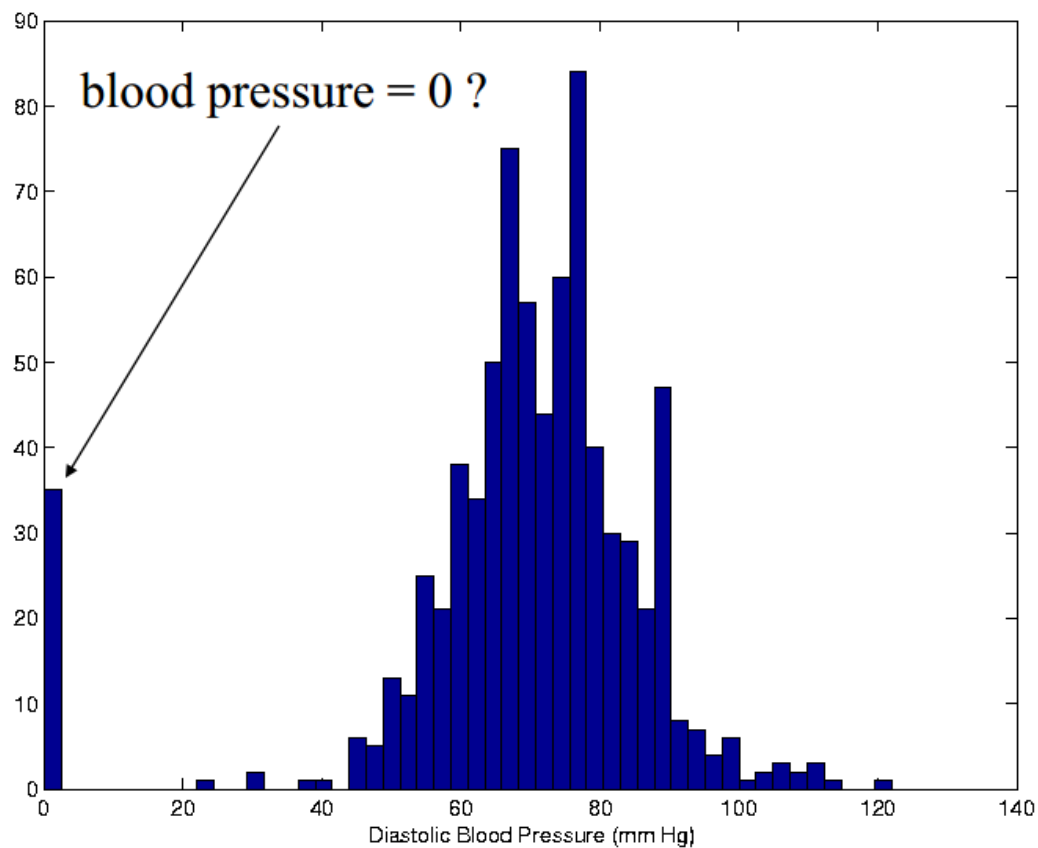
Histograms – Example

- Petal Width (10 and 20 bins, respectively)





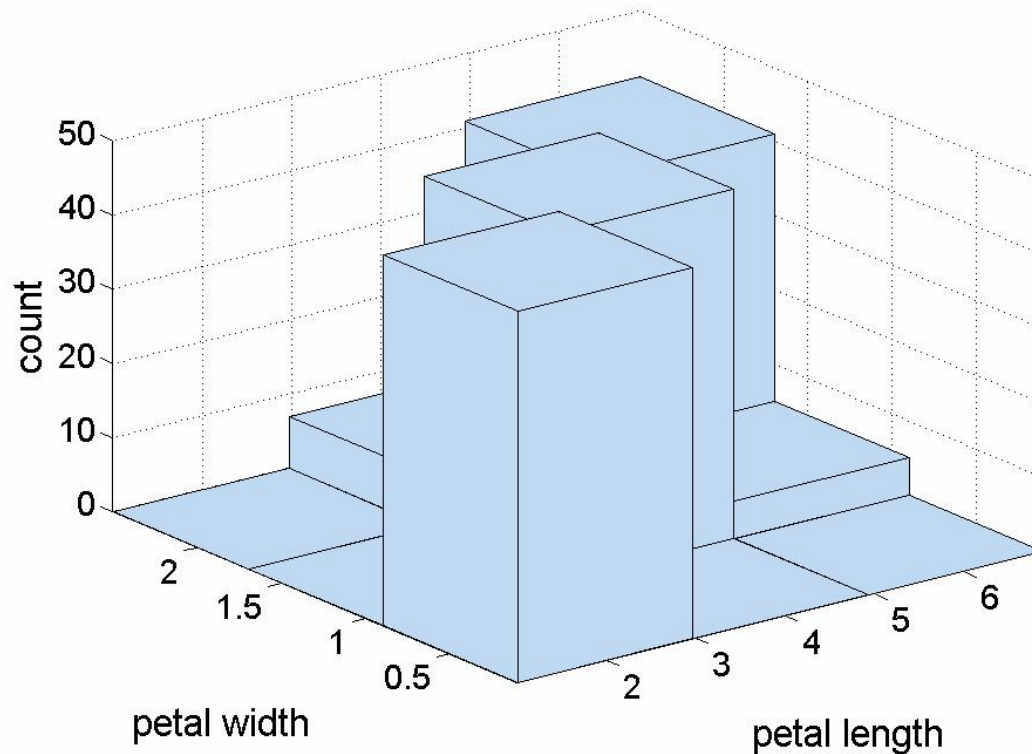
Hystogram and outliers





2-Dim Histograms

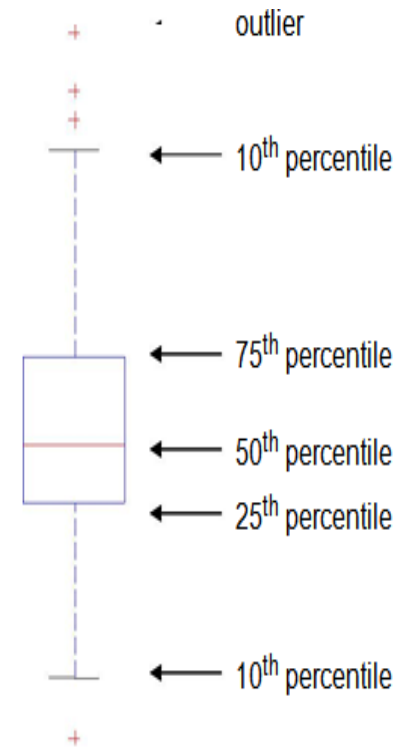
- Show the joint distribution of the values of two attributes





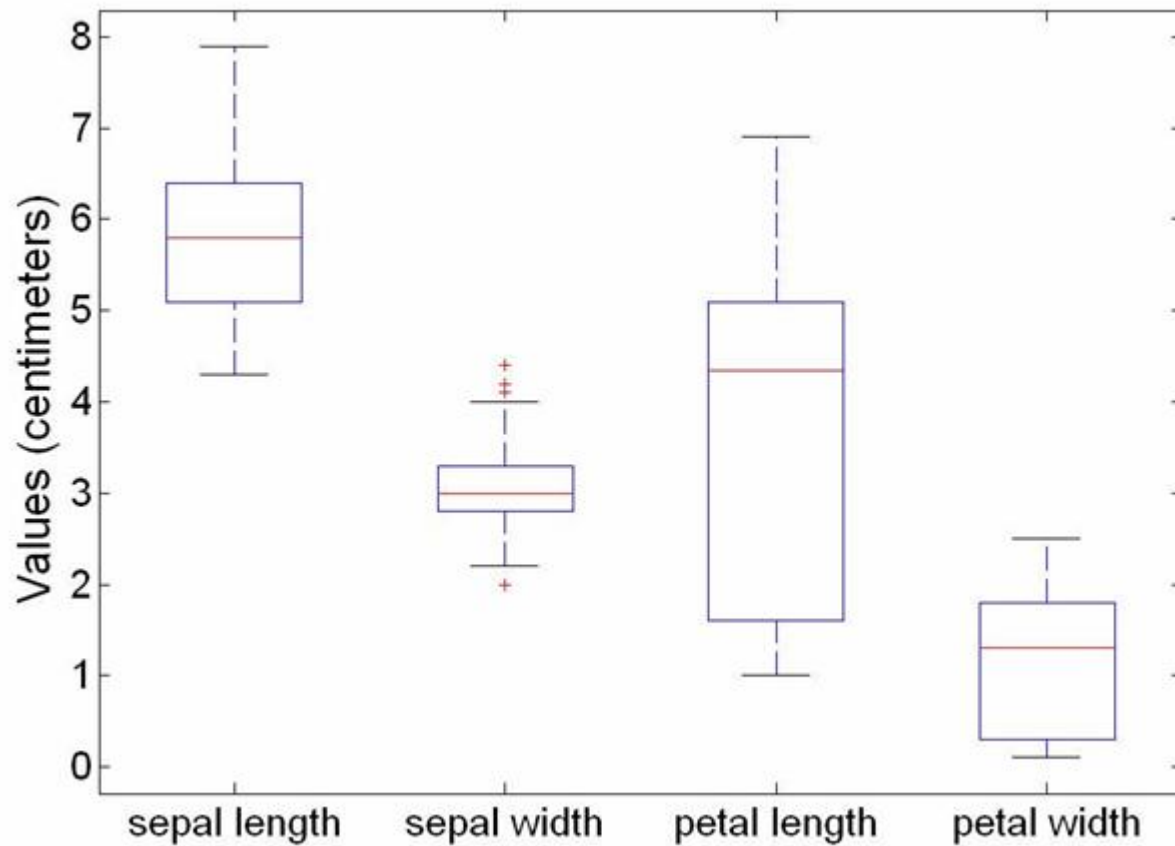
Visualization Techniques- BoxPlot

- Data is represented with a box
- The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
- The median is marked by a line within the box
- Whiskers: two lines outside the box extended to Minimum and Maximum
- Outliers: points beyond a specified outlier threshold, plotted individually





BoxPlots – Example



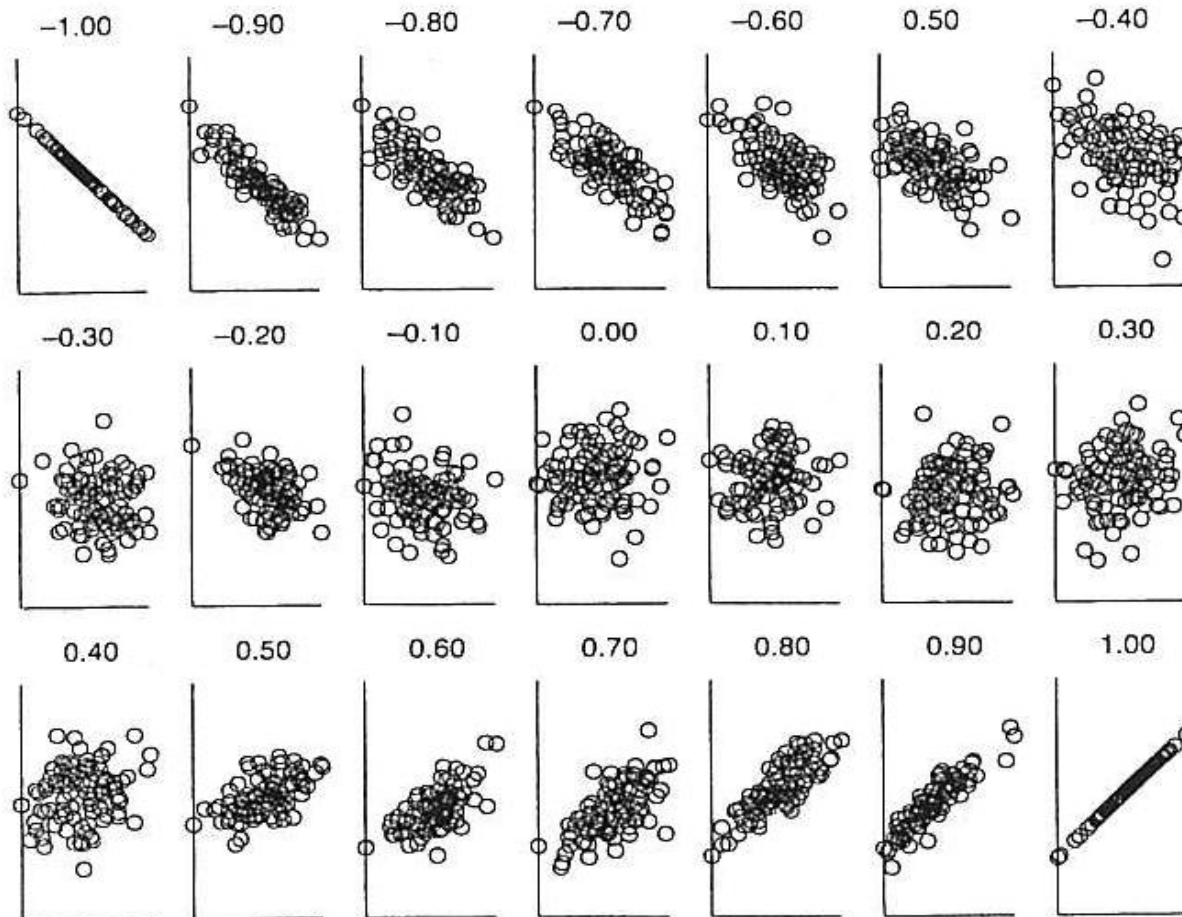


Visualization Techniques- Scatter Plots

- Used to discovery **linear correlation** between attributes
- Attributes values determine the position
- **Additional attributes** can be displayed by using the size, shape, and color of the markers that represent the objects
- *Arrays* of scatter plots can compactly summarize the relationships of several pairs of attributes
- The two-dimensional scatter plots are the most common, but we can have *three-dimensional* scatter plots



Correlations in scatter plots





Scatter Plots – Example

