SDSC 5001: Statistical Machine Learning I

Topic 2. Exploring Data

What Is Data?

- Collection of data objects and their attributes
- Object is also known as record, point, case, sample, entity, or instance
- An attribute is a property or characteristic of an object
  - Examples: age, height, weight, education of a person
  - Attribute is also known as variable, field, characteristic or feature
- A collection of attributes that describe an object

A Toy Model

ID	Refund	Marital Status	Taxable Income	···
1	Yes	Single	125K	···
2	No	Married	100K	···
3	No	Single	70K	···
4	Yes	Married	120K	···
5	No	Divorced	95K	···
⋮				

Types of Variables

- **Continuous variable**
  - Example: length, time, count, weight, height
- **Nominal (categorical) variable**
  - Example: race, sex, marital status, eye color
- **Ordinal variable**
  - Example: age group (children, youth, adults, seniors), letter grade, satisfaction rating (“dislike”, “neutral”, “like”)
- **Interval variable**
  - Example: temperature, salary range

Types of Data

- Data matrix
- Text data, image data, audio data, video data
- Transactions data
- Graph data

Data Matrix

- If data objects have the same fixed set of numeric variables, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct variable.
- Such a data set can be represented by an $n \times p$ matrix, where there are n rows, one for each object, and p columns, one for each variable.

Patient index	Sex	Age	Smoker (Y/N)	Alcohol (Y/N)
1	F	28	N	Y
2	M	35	N	N
3	M	60	Y	Y

Text Data

- Each textual document becomes a vector of terms
- Each term corresponds to a word or phrase
- The element of the vector is the number of times the corresponding term appears in the textual document.
- Example:
  - Doc 1: I love dogs.
  - Doc 2: I hate dogs and knitting.
  - Doc 3: Knitting is my hobby and passion.

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1	0	0	0	0	0	0	0
Doc 2	1	0	1	1	1	1	0	0	0	0
Doc 3	0	0	0	0	1	1	1	1	1	1

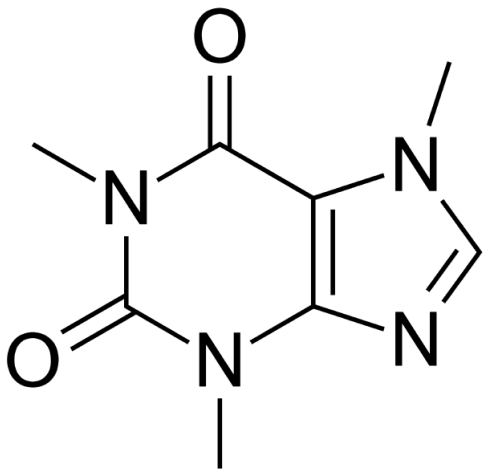
Transaction Data

- A special type of record data
- Each record (transaction) involves a set of items
- Example: In a grocery store, the set of products purchased by a customer during one shopping trip constitute a transaction

ID	Products
1	Masks, Bread, Coke, Milk
2	Masks, Beer, Bread, Diapers
3	Masks, Beer, Diapers
4	Masks, Beer, Bread, Milk
5	Masks, Beer, Coke, Diapers, Milk

Graph Data

- Social network
- Molecular structures

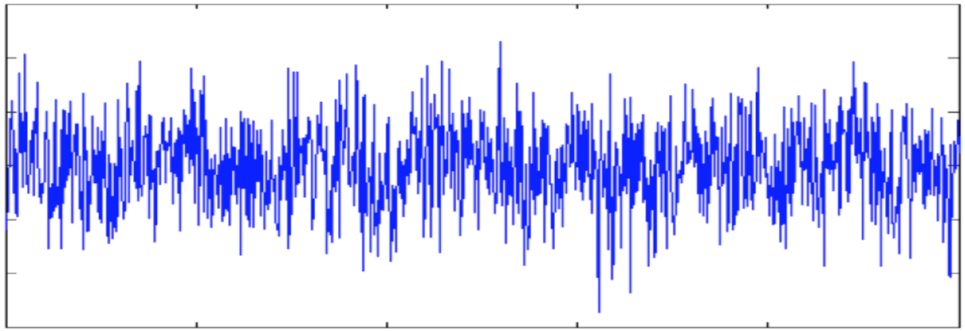


Data Quality

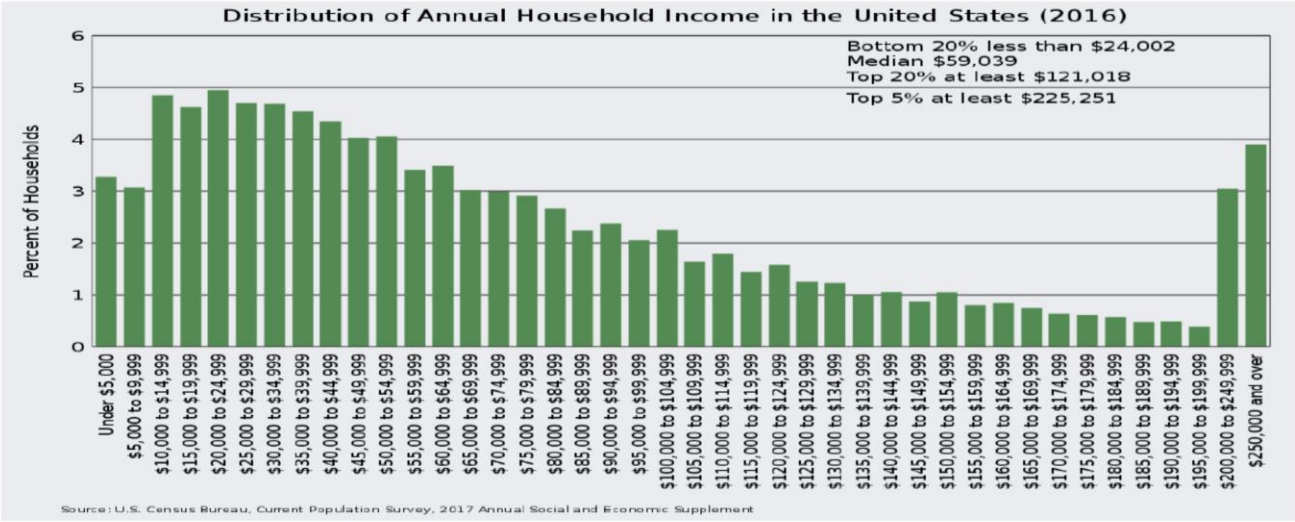
- **Quality of data** is a very important factor for success of subsequent analysis.
- **Possible data quality issues**
  - Noise and outliers
  - Missing values
  - Sampling bias

Noise and Outliers

- **Noise** refers to perturbation of original values.



- **Outliers** are observations that are considerably different from others.



Missing Values

- **Reasons for missing values**
  - Information is not collected (e.g., people decline to give their age and weight)
  - Variables may not be applicable to all cases (e.g., annual income is not applicable to children)
- **Handling missing values**
  - Eliminate all data objects with missing value
  - Impute missing values
  - Incorporate partial information of the missing values in data analysis

Sampling Bias

- Sample distortion arises from a mismatch between the random sample and the population of interest.
- **Causes of sampling bias**
  - Convenience sampling
  - Class imbalance (e.g., anomaly detection, disease prediction)
- **Avoiding and correcting** sampling bias

What Is Data Exploration?

- Preliminary exploration of the data to better understand its characteristics
  - Key motivations of data exploration include
    - Helping select the right tool for preprocessing or formal analysis
    - Making use of human's abilities to recognize patterns
  - Exploratory data analysis (EDA)
    - Created by statistician John Turkey; see his seminal book on EDA

Data Exploration Techniques

- 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
- We will focus on
  - Summary statistics
  - Visualization

Summary Statistics

- Summary statistics are numbers that summarize properties of the data.
- Examples
  - Frequency: the percentage of time a given value appears in the data set
  - Mode: the value that appears most frequently
  - Location: mean, median, trimmed mean, percentile
  - Spread: range, variance, standard deviation
  - Skewness

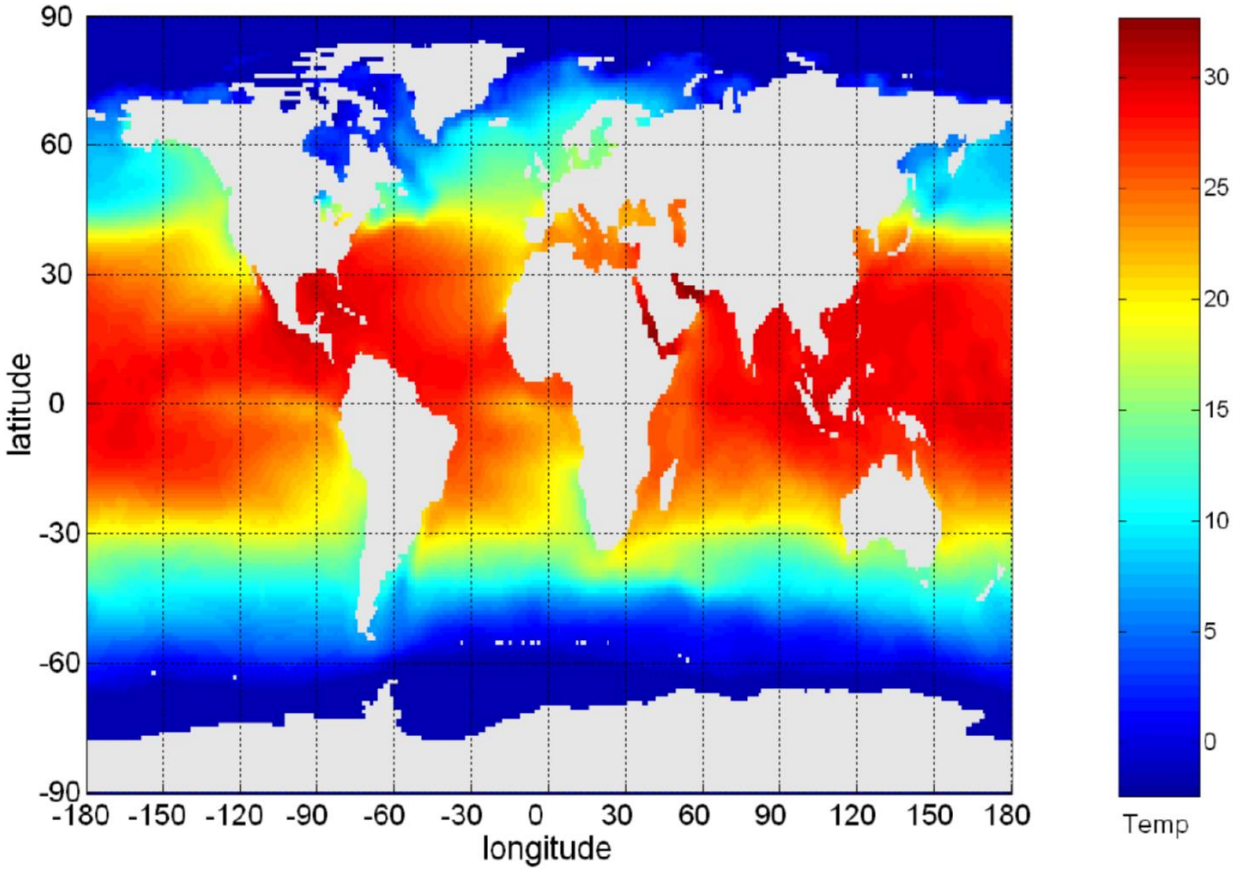
Visualization

- Visualization is the conversion of data into a visual format so that the characteristics of the data and the relationships among data objects or variables 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.
  - It can detect general patterns and trends.
  - It can detect outliers and unusual patterns.

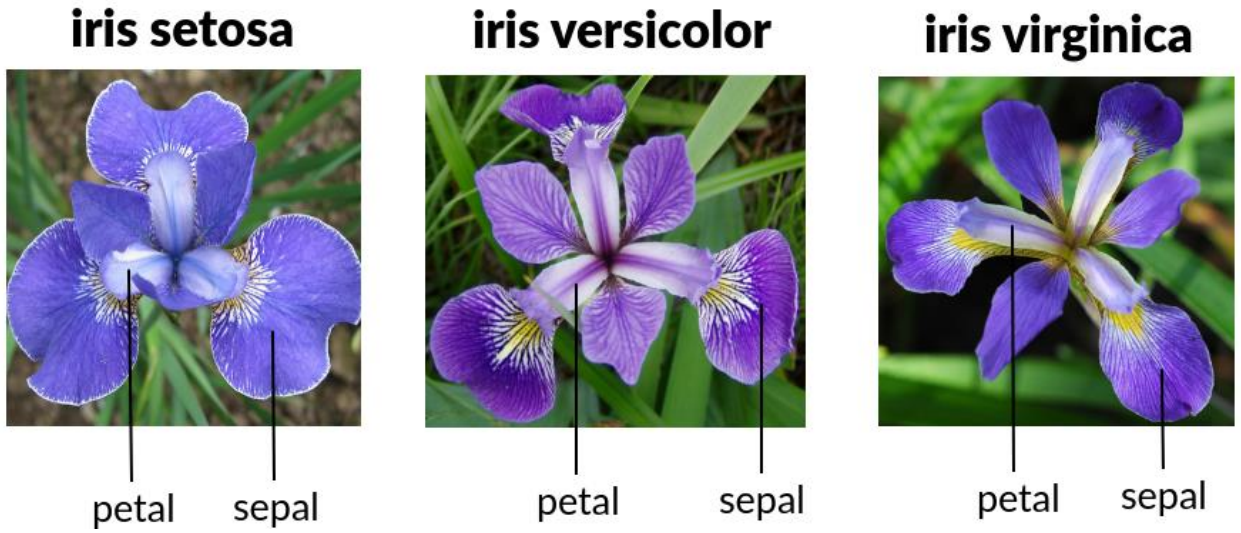
Sea Surface Temperature

- Sea surface temperature for July 1982
- Tens of thousands of data points are summarized in one single figure

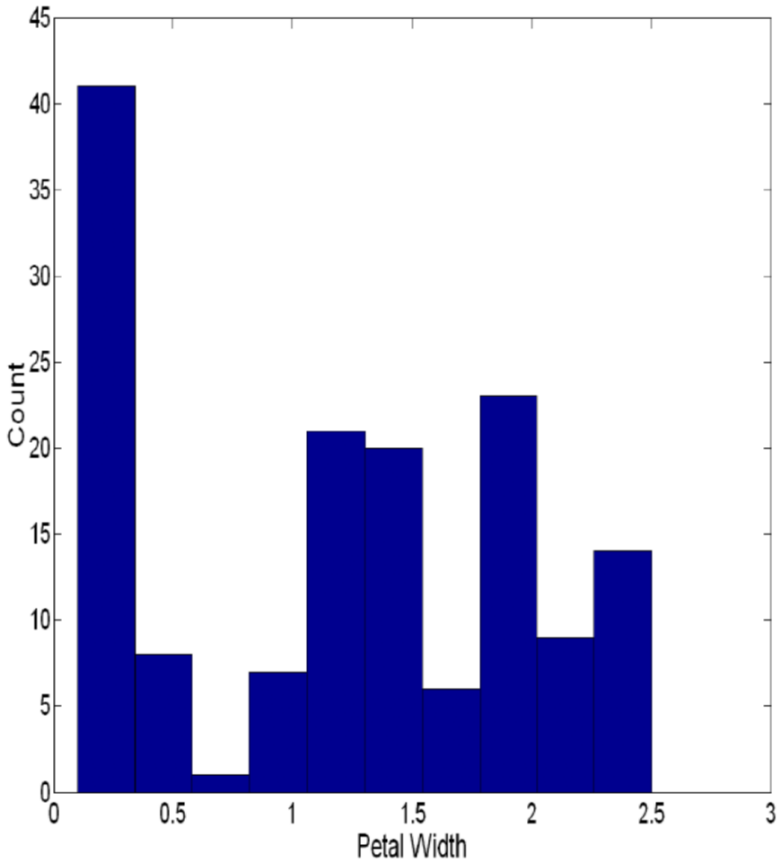


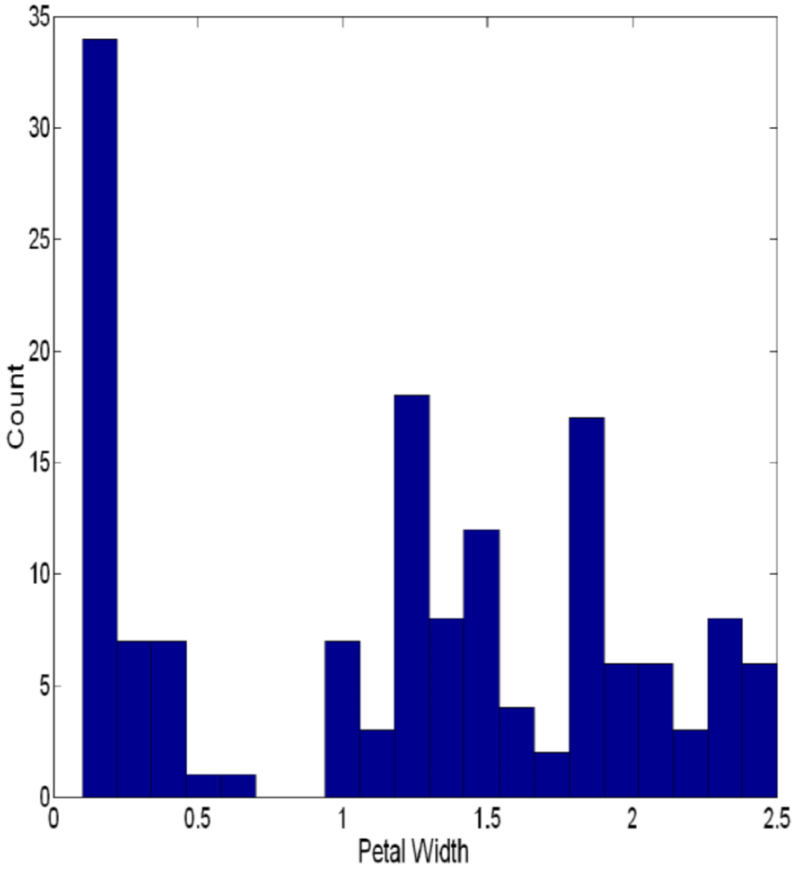
Iris Data

- A classical dataset
- Three iris types (classes): Setosa, Versicolor, Virginica
- Four explanatory variables
  - Sepal width and length
  - Petal width and length
- Sample size: 50 samples for each type

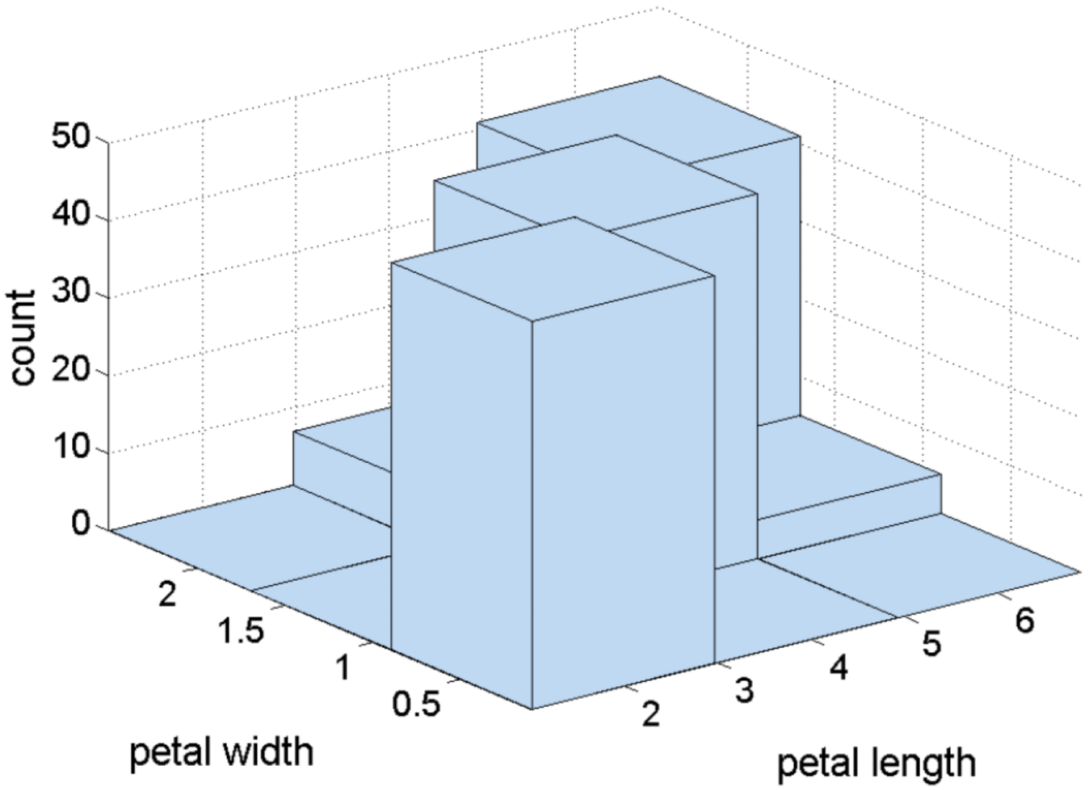


Histogram

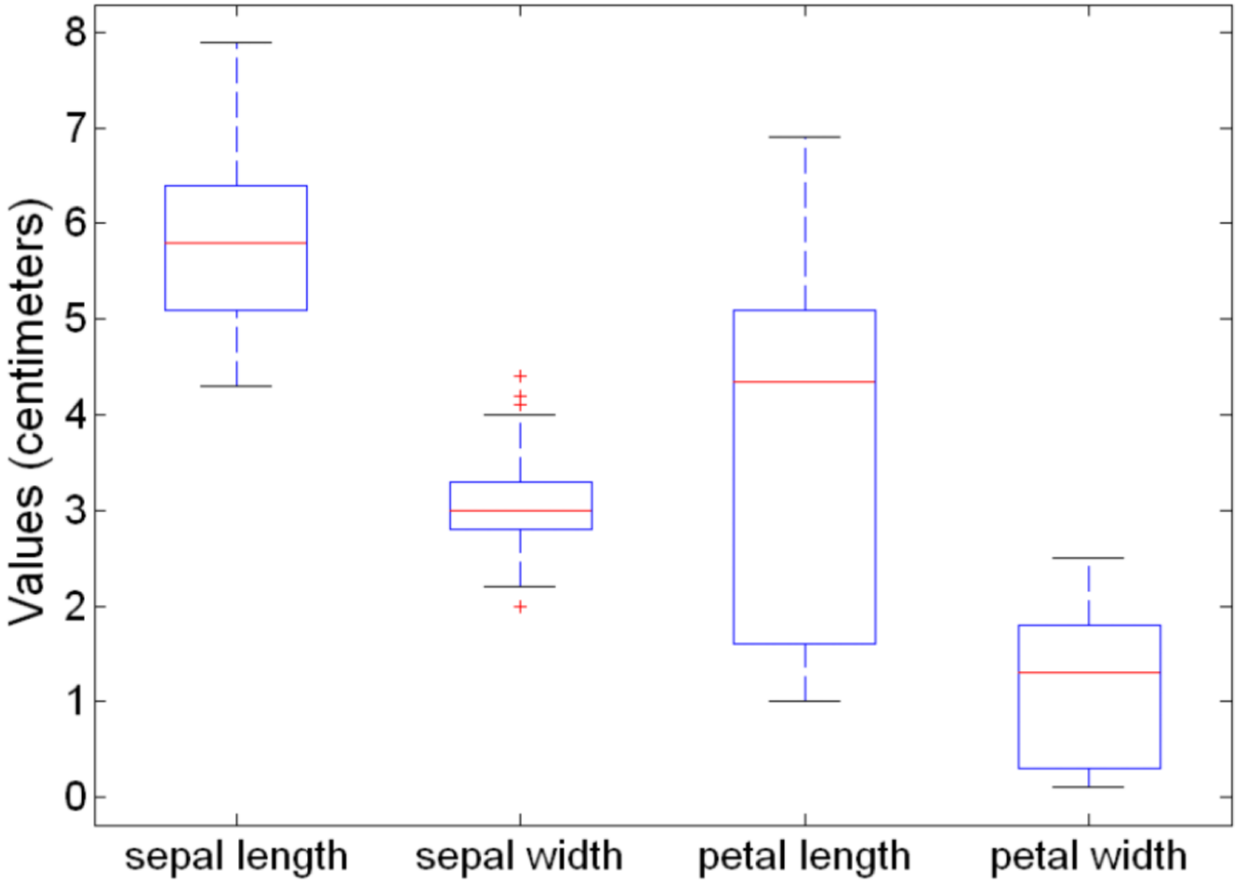

Count
Petal Width
0
5
10
15
20
25
30
35
40
45
0
0.5
1
1.5
2
2.5
3


Count
Petal Width
0
5
10
15
20
25
30
35
0
0.5
1
1.5
2
2.5

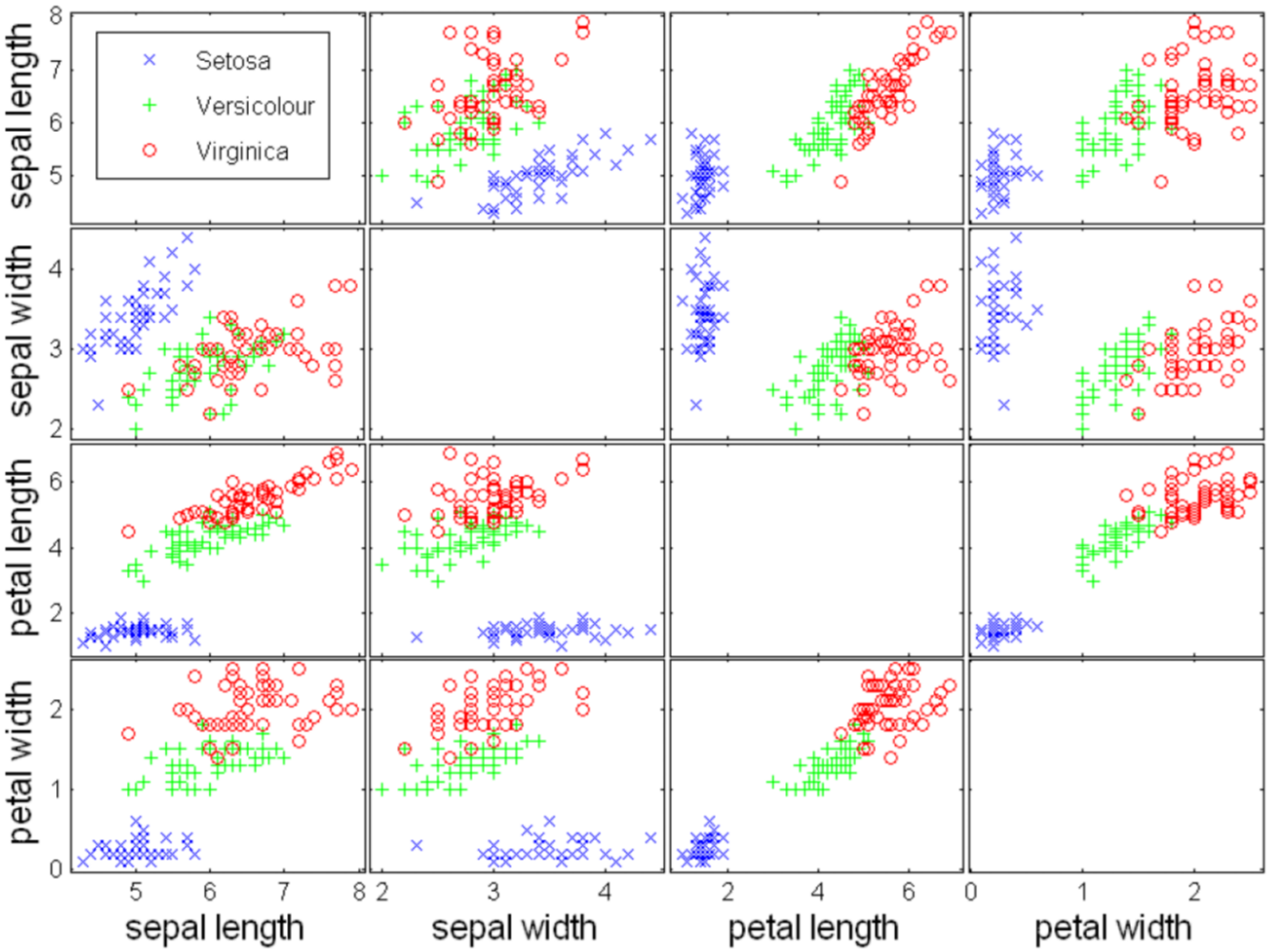
2D Histogram



Boxplot


Values (centimeters)
0
1
2
3
4
5
6
7
8
sepal length
sepal width
petal length
petal width

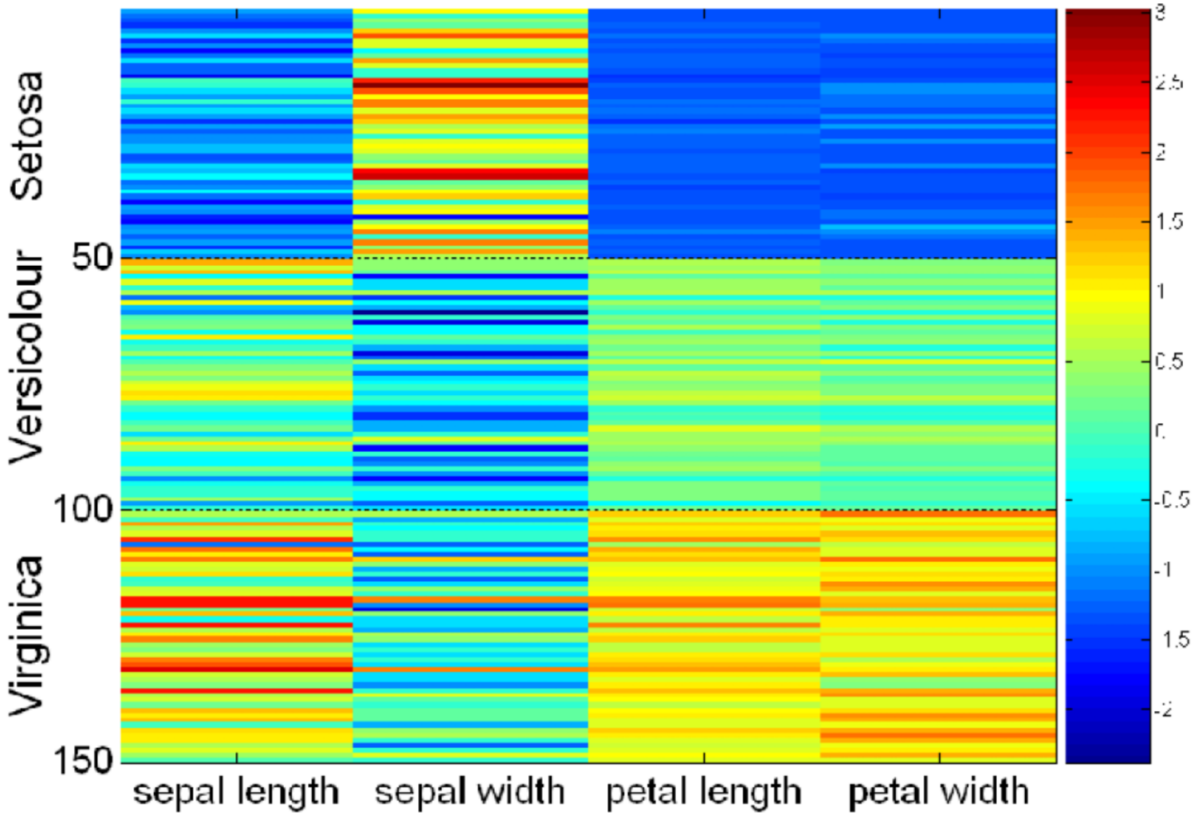
Scatter Plot



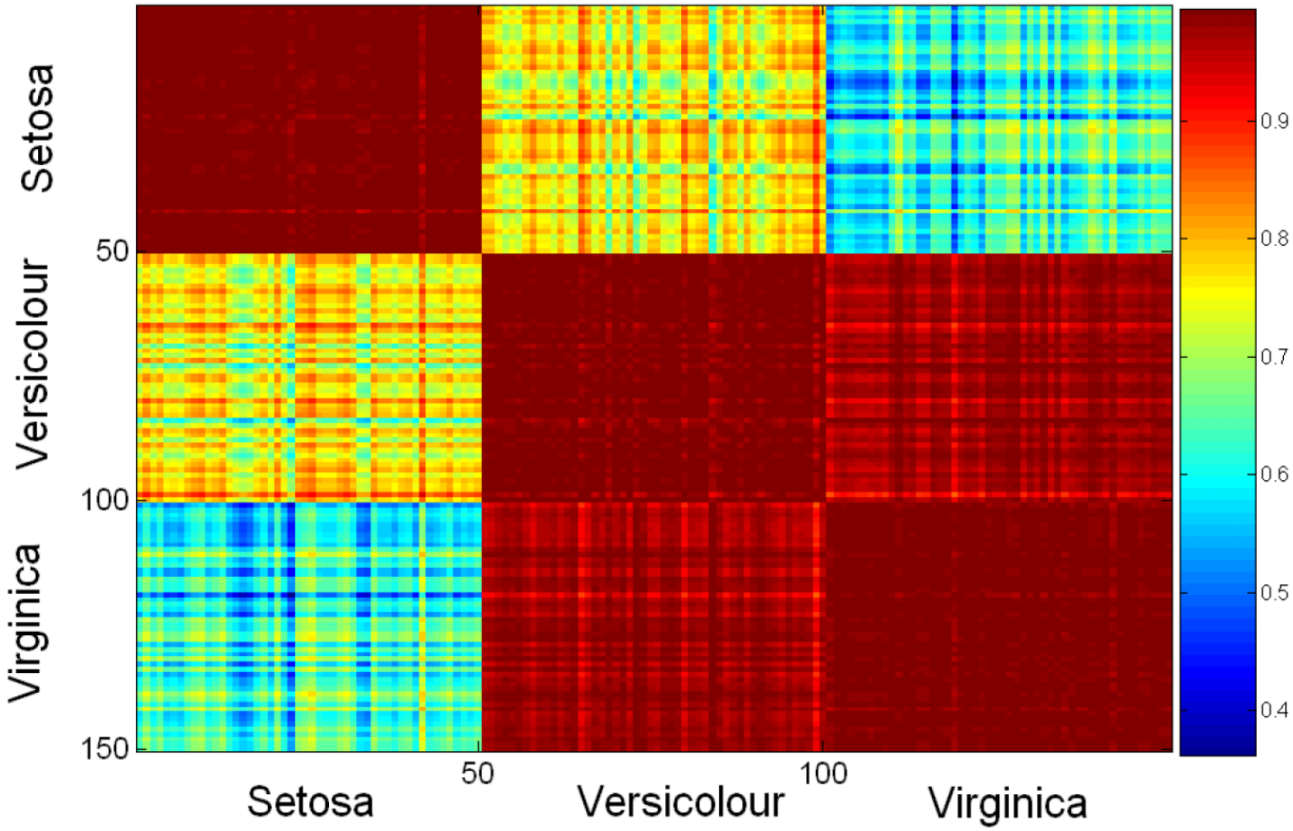
Matrix Plot

- Plot the data matrix
- This can be useful when objects are sorted according to class.
- Typically, the variables are normalized to prevent one variable from dominating the plot.
- Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects.

Iris Matrix Plot



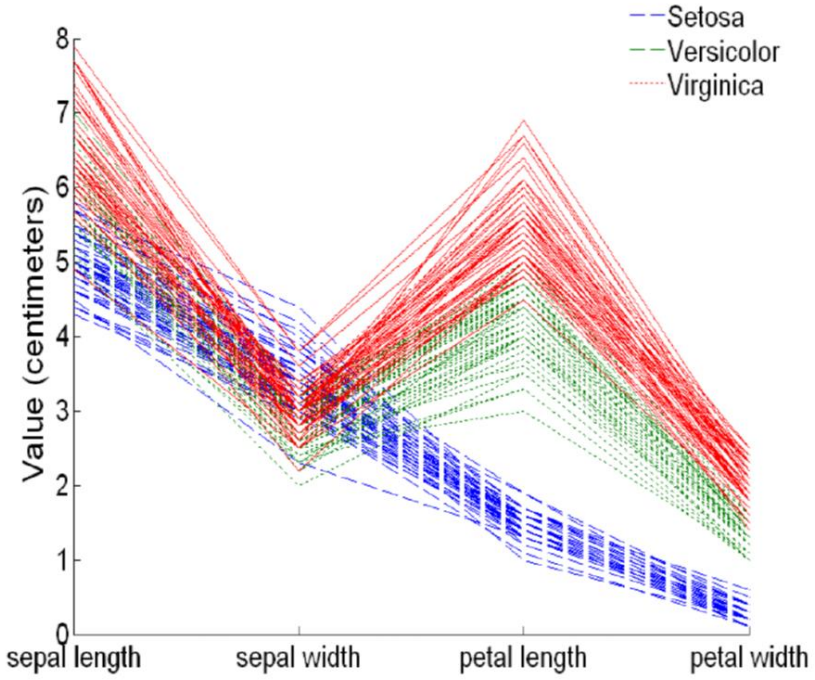
Iris Similarity Matrix

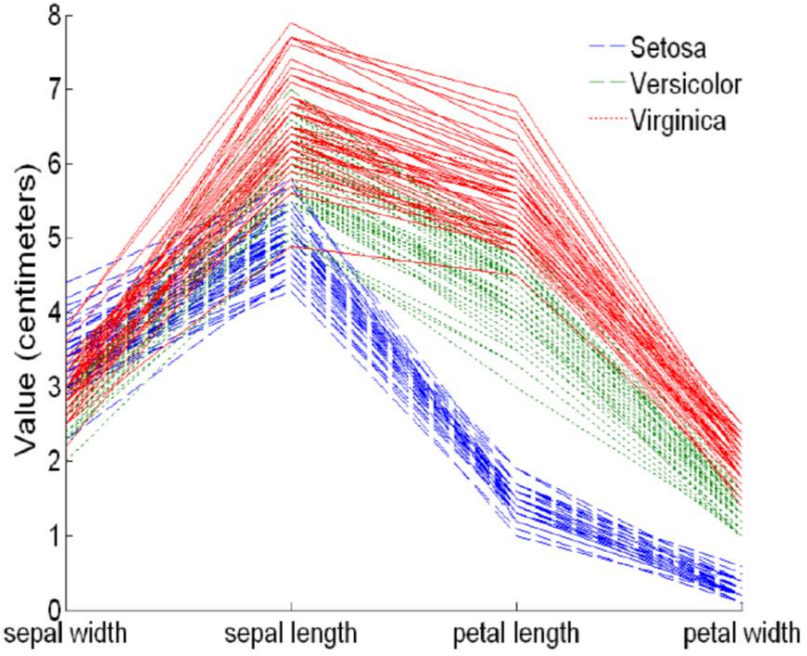


Parallel Coordinates Plot

- Use a set of parallel axes
- The variable 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 by a line.
- Often, the lines representing a distinct class of objects group together, at least for some variables.
- Ordering of variables can be important.

Example

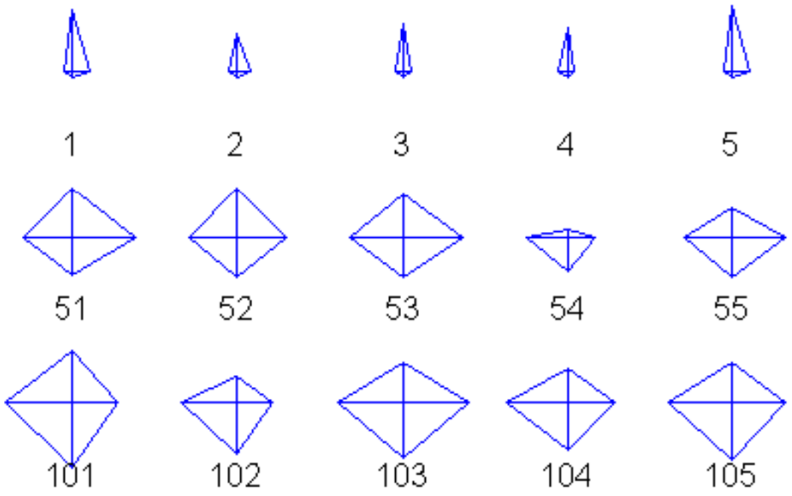

Setosa
Versicolor
Virginica
Value (centimeters)
0
1
2
3
4
5
6
7
8
sepal length
sepal width
petal length
petal width

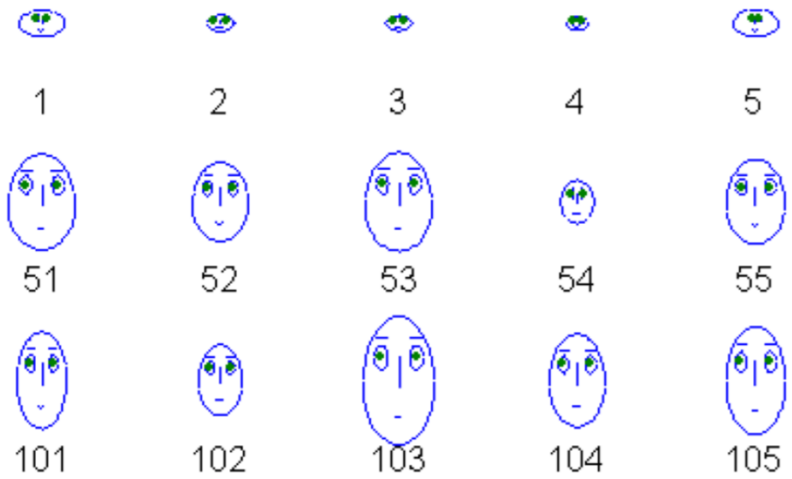

Setosa
Versicolor
Virginica
Value (centimeters)
0
1
2
3
4
5
6
7
8
sepal width
sepal length
petal length
petal width

Other Visualization Techniques

- Star plots
  - Axes radiate from a central point
  - Each object becomes a polygon
- Chernoff faces
  - Associate each variable with a characteristic of a face
  - The values of each variable determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face

Example


1
2
3
4
5
51
52
53
54
55
101
102
103
104
105


1
2
3
4
5
51
52
53
54
55
101
102
103
104
105