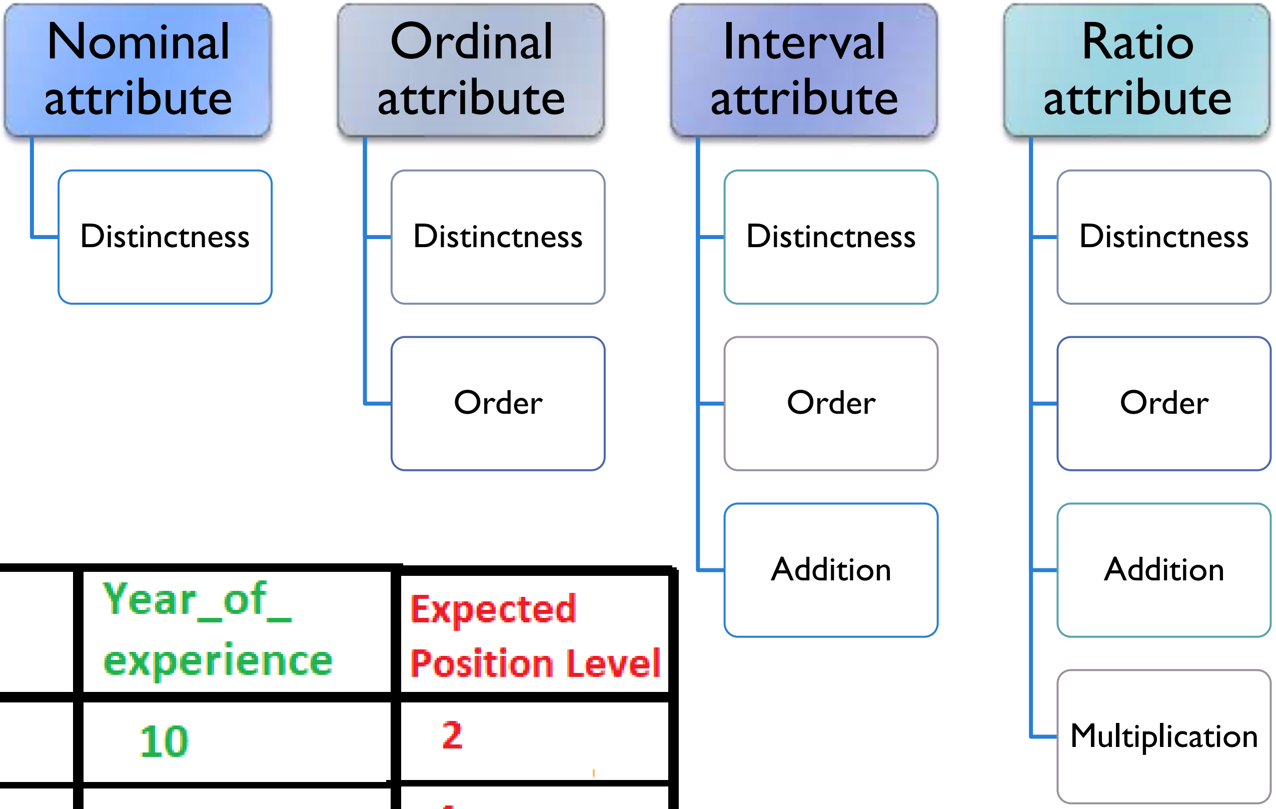
QUICK REVIEW



Nominal attribute
- Distinctness

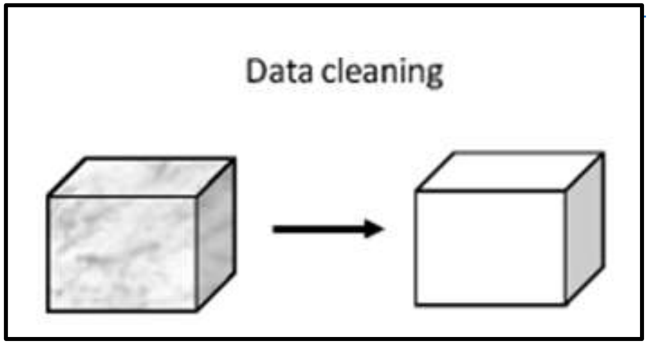
Ordinal attribute
- Distinctness
- Order

Interval attribute
- Distinctness
- Order
- Addition

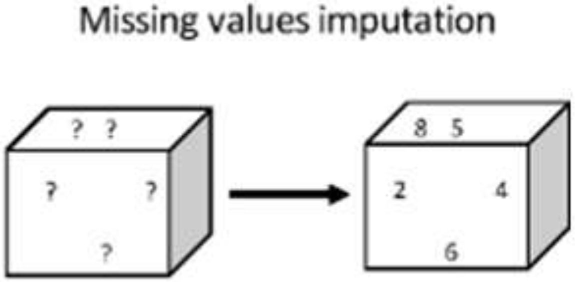
Ratio attribute
- Distinctness
- Order
- Addition
- Multiplication

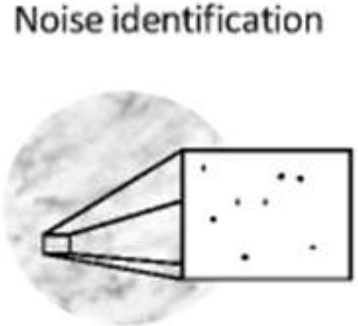
person_name	Salary	Year_of_ experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

Forms of data preprocessing

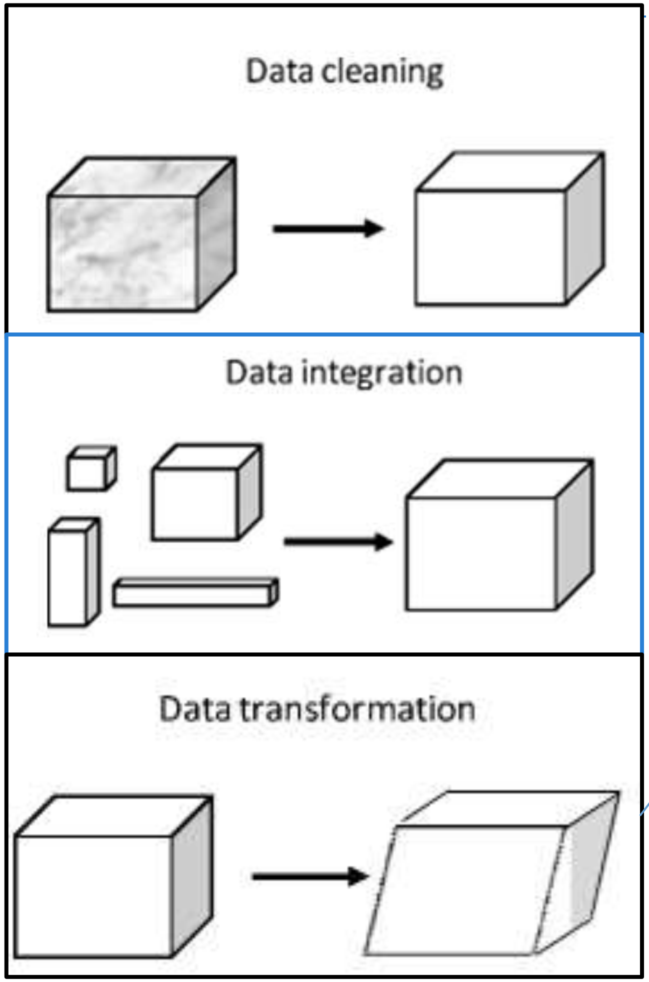


- Fill in missing values
- Smooth noisy data
- Remove outliers
- Resolve inconsistencies

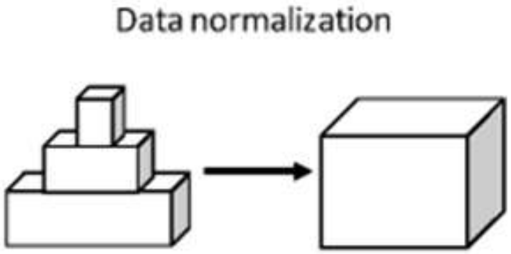




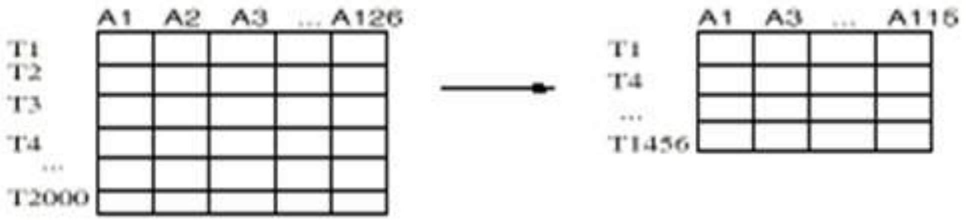
Forms of data preprocessing



- Fill in missing values
- Smooth noisy data
- Remove outliers
- Resolve inconsistencies

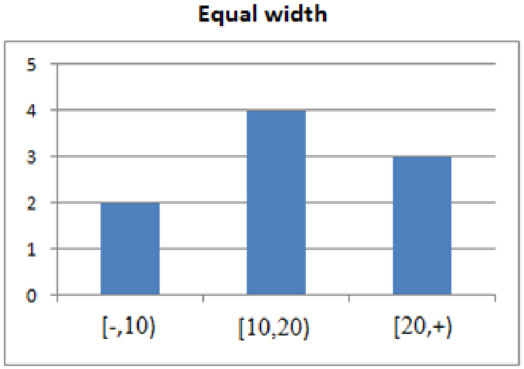


Data Reduction

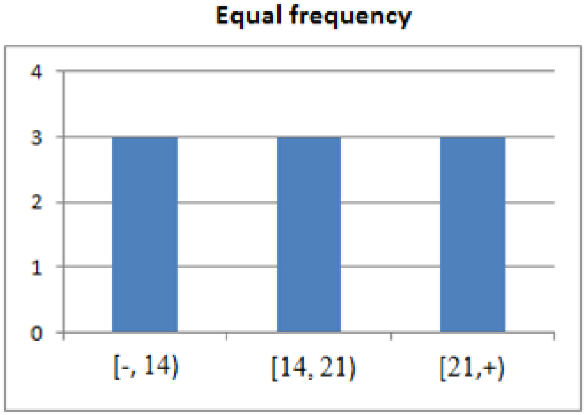


Binning Example

- Attribute values (for an attribute age):
  - 0, 4, 12, 16, 16, 18, 24, 26, 28 Sorted
- Equi-width binning – for bin width of 10:
  - Bin 1: 0, 4 [-,10) bin
  - Bin 2: 12, 16, 16, 18 [10,20) bin
  - Bin 3: 24, 26, 28 [20,+) bin
  - – denote negative infinity, + positive infinity



- Equi-frequency binning – for bin density of 3
  - Bin 1: 0, 4, 12 [-, 14) bin
  - Bin 2: 16, 16, 18 [14, 21) bin
  - Bin 3: 24, 26, 28 [21,+] bin



Binning Methods for Data Smoothing

* Sorted data for price: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**

* Partition into Equi-depth bins:

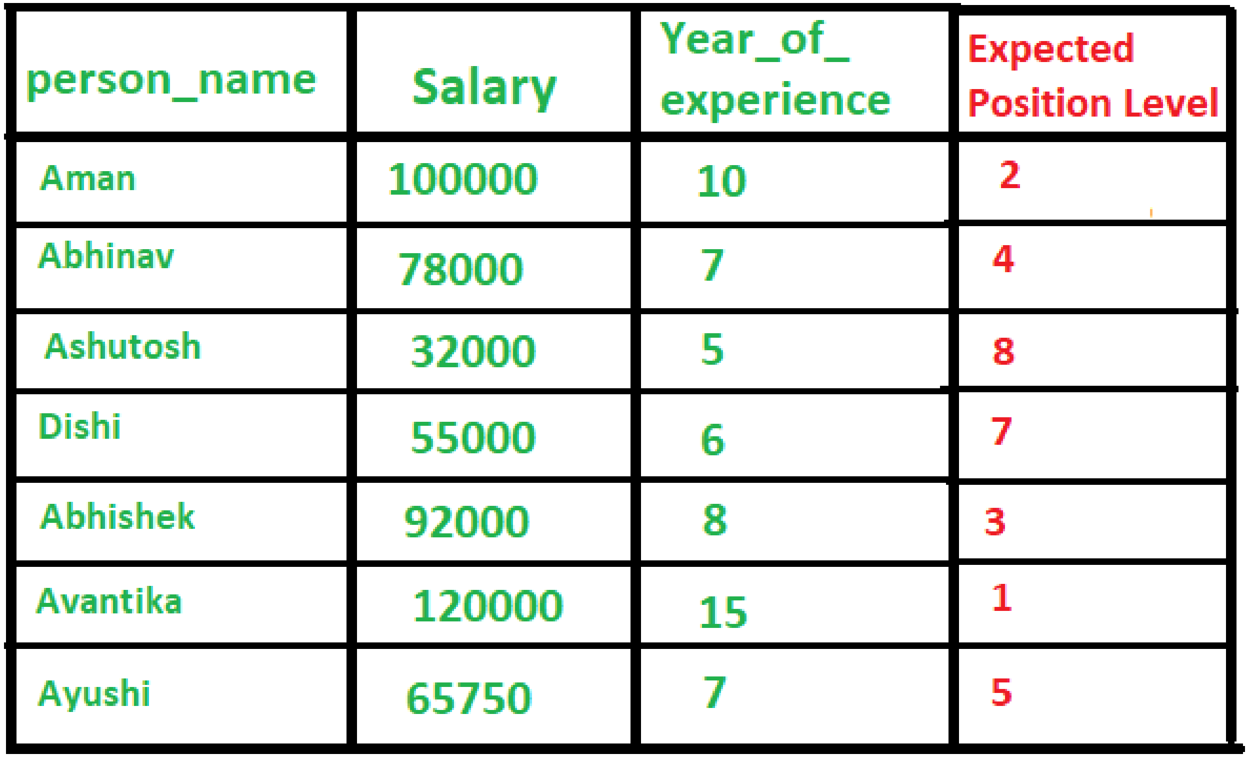
Equi-depth bins:
Bin 1: 4, 8, 9, 15
Bin 2: 21, 21, 24, 25
Bin 3: 26, 28, 29, 34

Smoothing by bin means:
Bin 1: Bin 1: 9, 9, 9, 9
Bin 2: 23, 23, 23, 23
Bin 3: 29, 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 4, 15
Bin 2: 21, 21, 25, 25
Bin 3: 26, 26, 26, 34

Data Transformation

Transform or consolidate data into forms appropriate for mining



person_name	Salary	Year_of_experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

Normalization: scaled to fall within a small, specified range

Data Transformation: Normalization

- An attribute values are scaled to fall within a small, specified range , such as 0.0 to 1.0

- **Min-Max normalization**
  - performs a linear transformation on the original data.

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- **Example:** Let min and max values for the attribute *income* are $12,000 and $98,000, respectively.
- Map *income* to the range [0.0;1.0].

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716.$$

Data Transformation: Normalization

- **z-score normalization(or *zero-mean normalization*)**
  - An attribute A, values are normalized based on the mean and standard deviation of *A*.

$$\boxed{v' = \frac{v - mean_A}{stand\_dev_A}}$$

- **Example:** Let mean= 54,000 and standard deviation=16,000 for the attribute *income*
- With z-score normalization, a value of $73,600 for *income* is transformed to

$$\boxed{\frac{73{,}600-54{,}000}{16{,}000} = 1.225.}$$

Data Reduction

- Warehouse may store terabytes of data
- Complex data analysis/mining may take a very long time to run on the complete data set

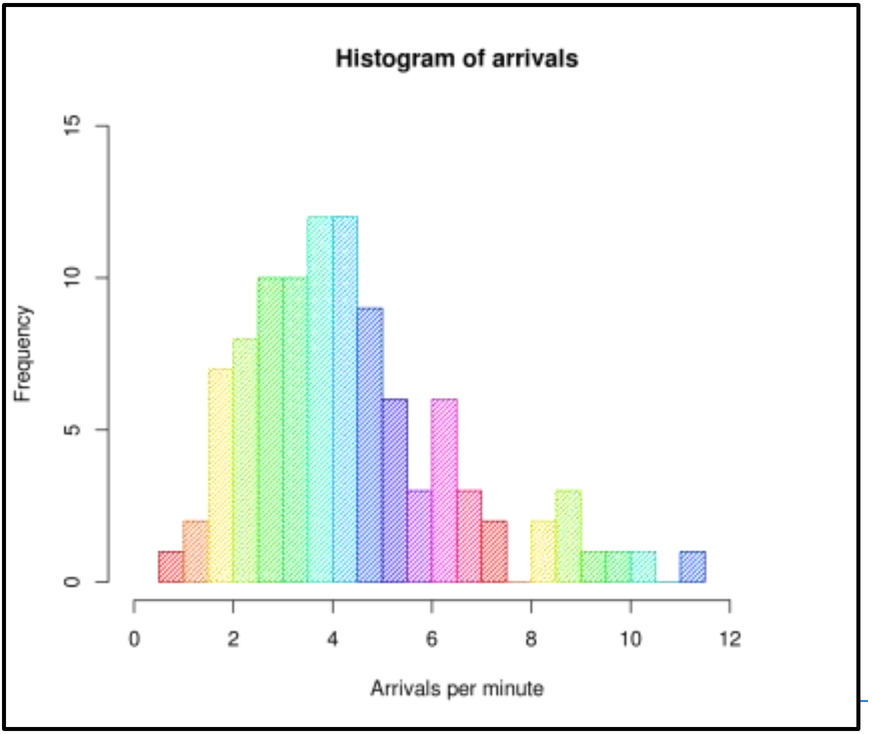
- **Data reduction**
  - Obtains a reduced representation of the data set that is much smaller in volume
  - but produces the same (or almost the same) analytical results

Data Reduction Strategies

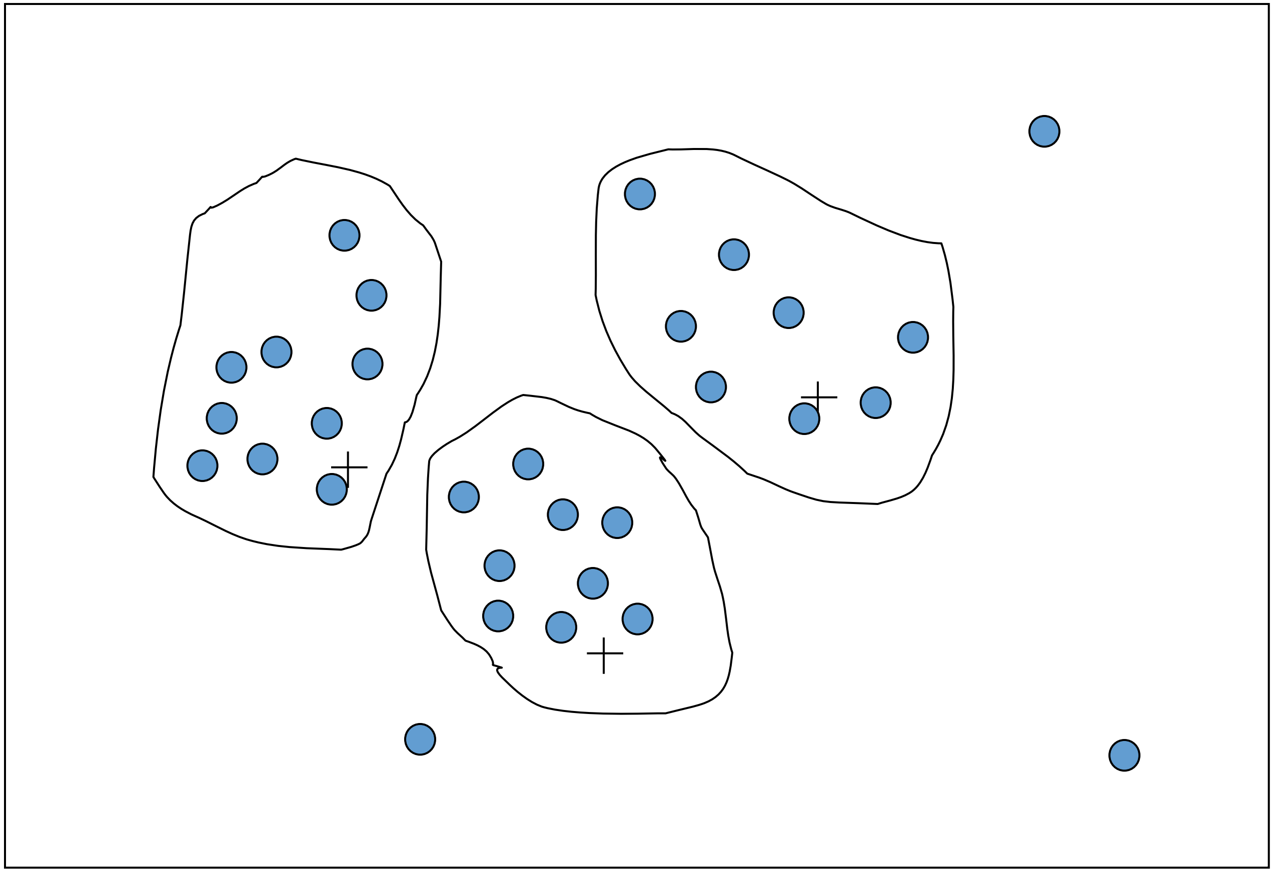
- **Dimensionality reduction**
- **Numerosity reduction**
  - data is replaced or estimated by alternative smaller data representations
    - Histogram
    - Sampling
    - Clustering
- **Discretization and concept hierarchy generation**
  - replace raw data values for attributes by ranges or higher conceptual levels

Numerosity reduction -Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Same as Binning
- Can be constructed optimally in one dimension using dynamic programming



Numerosity reduction - Cluster Analysis



Partition data into clusters, and store cluster representation only

Can be very effective if data is in form of clusters

Numerosity reduction - Sampling

Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

Example: What is the average height of a person in Pakistan?

We cannot measure the height of everybody

Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Example: We have 1M documents. How many has at least 100 words in common?

- Computing number of common words for all pairs requires **10^{12}** comparisons

Example: What fraction of tweets in a year contain the word "Lahore"?

- 300M tweets per day, if 100 characters on average, 86.5TB to store all tweets

Types of Sampling

Simple Random Sampling

- There is an equal probability of selecting any particular item

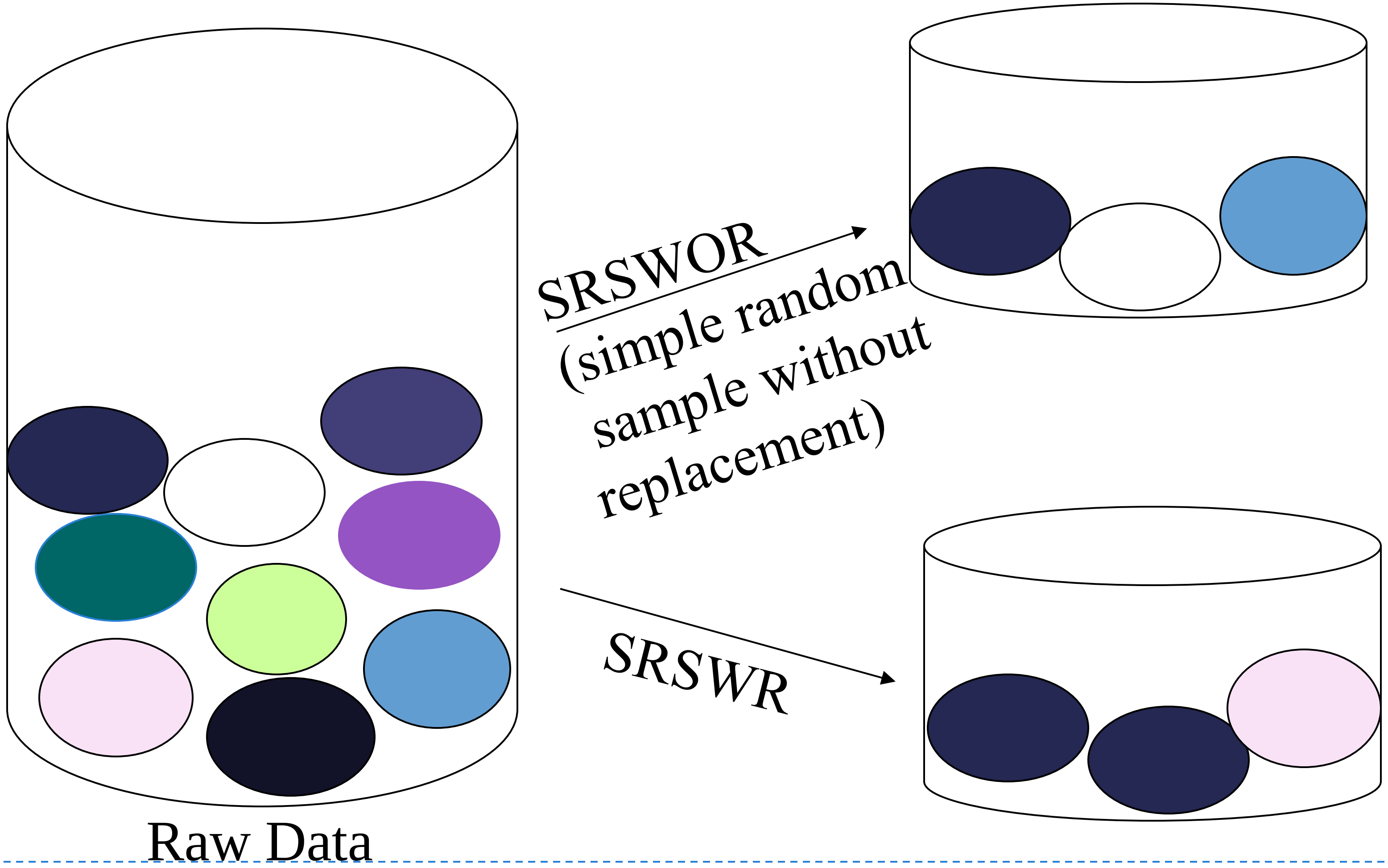
Sampling without replacement

- As each item is selected, it is removed from the population

Sampling with replacement

- Objects are not removed from the population as they are selected for the sample.
- In sampling with replacement, the same object can be picked up more than once.
- This makes analytical computation of probabilities easier

Sampling



SRSWOR
(simple random sample without replacement)
SRSWR
Raw Data

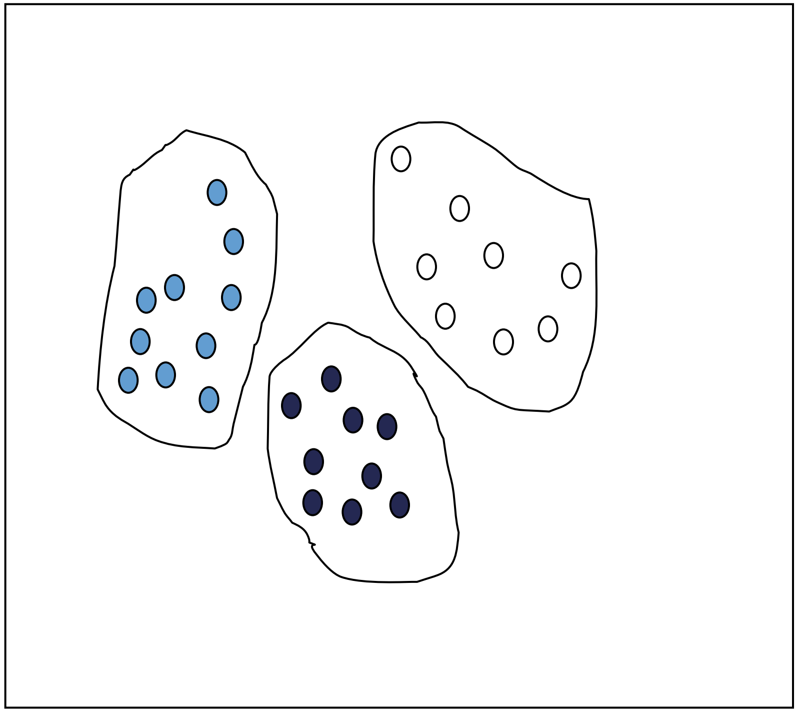
Types of Sampling

- **Stratified sampling**
  - Split the data into several **groups**; then draw random samples from each group.
  - Ensures that both groups are represented.

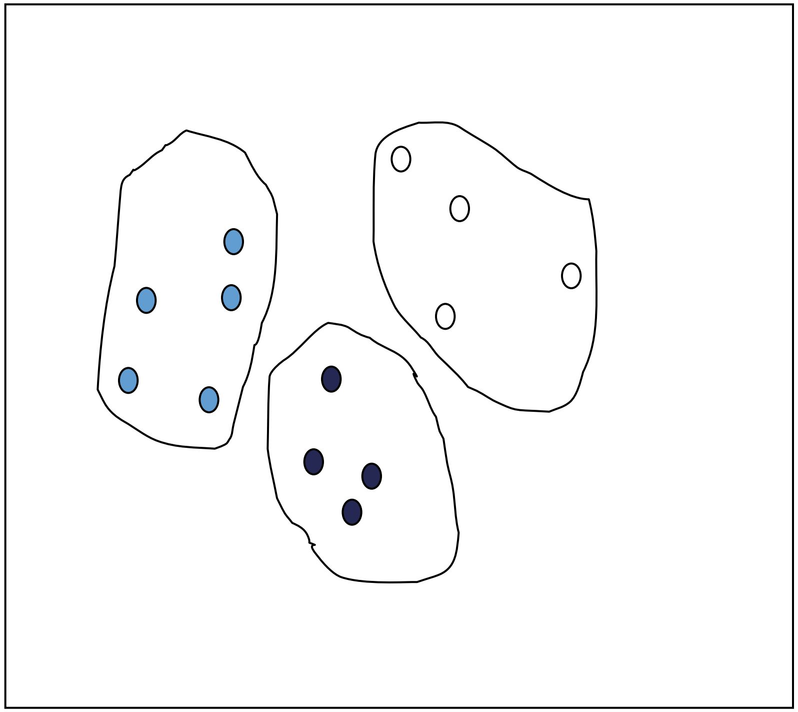
  - **Example** Find difference between legitimate and fraudulent credit card transactions.
  - 0.1% of transactions are fraudulent. What happens if we select 1000 transactions at random?
    - We get 1 fraudulent transaction (in expectation). Not enough to draw any conclusions.
    - Solution: sample 1000 legitimate and 1000 fraudulent transactions

Sampling

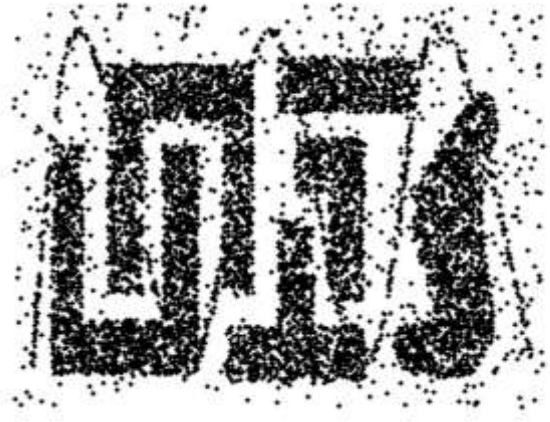
Raw Data



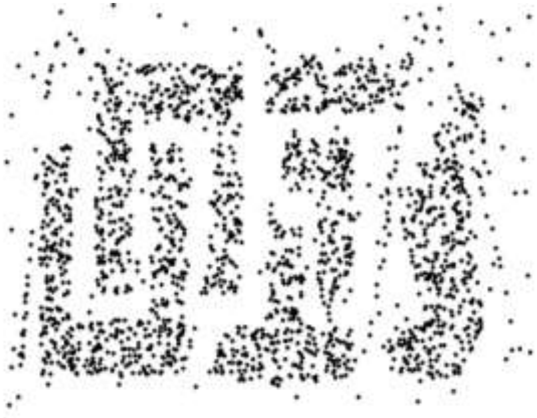
Cluster/Stratified Sample



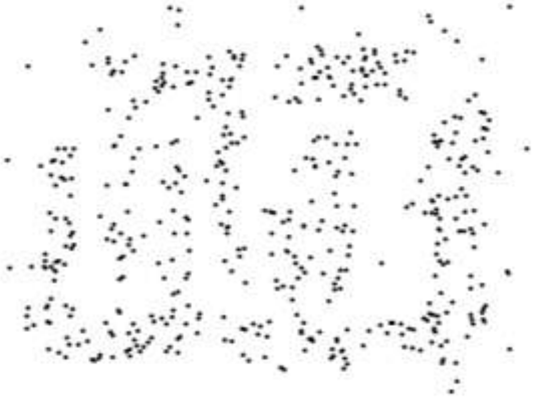
Sample Size



8000 points
Points



2000 Points



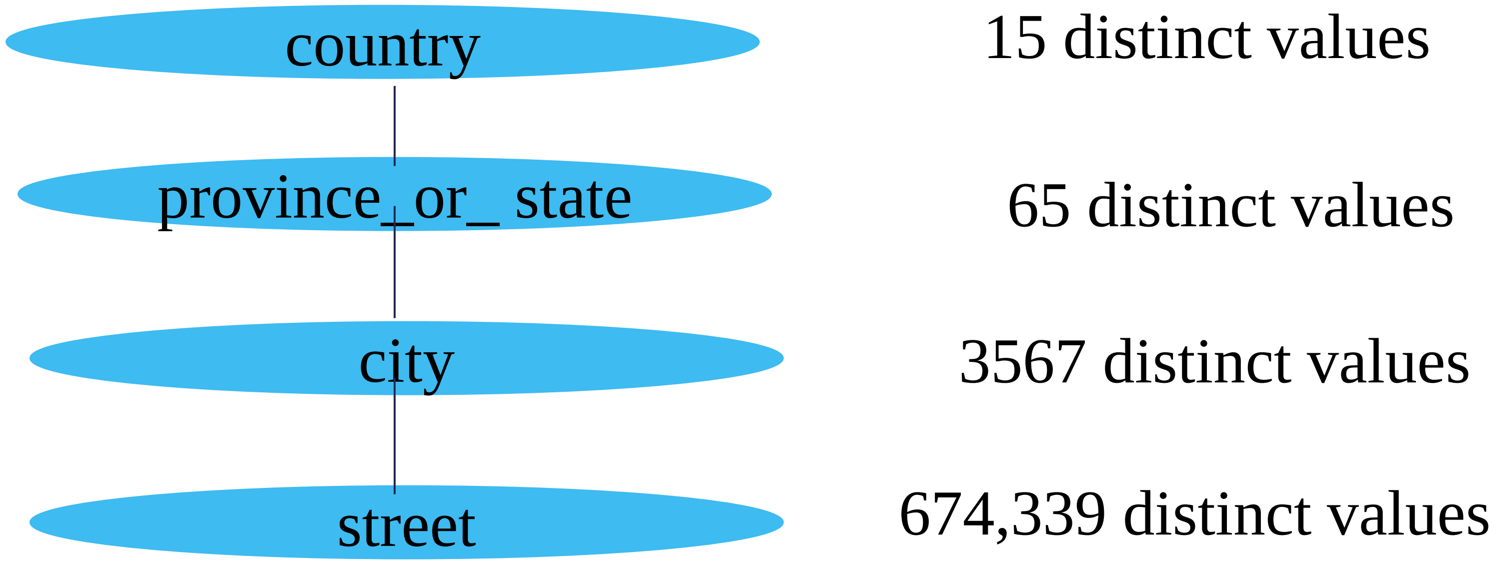
500

Concept hierarchy

- Concept hierarchy
  - Reduce the data by replacing low level concepts by higher level concepts
  - Replace numeric values for the attribute age by higher level concepts such as
    - young, middle-aged, or senior

Automatic Concept Hierarchy Generation

- Some concept hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the given data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy



Dimensionality Reduction

- Purpose
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Auto encoders
  - Others: supervised and non-linear techniques

Feature selection

- Another way to reduce dimensionality of data

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features
    - such that the probability distribution of different classes given the values of the selected features is as close to the original distribution given the values of all features

Feature Subset Selection

- **Redundant features**
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- **Irrelevant features**
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Subset Selection

Brute-force approch

- Try all possible feature subsets as input to data mining algorithm

Embedded approaches

- Feature selection occurs naturally as part of the data mining algorithm

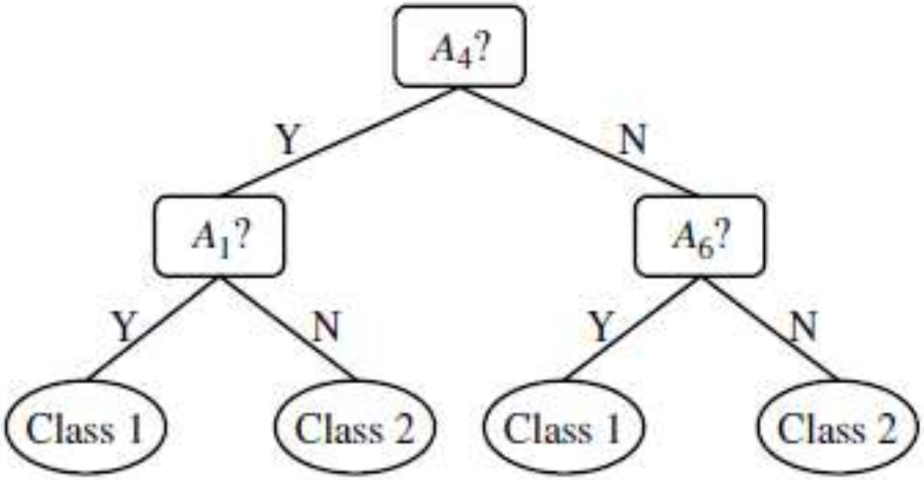
Filter approaches

- Features are selected before data mining algorithm is run

Wrapper approaches

- Use the data mining and machine learning algorithm as a black box to find best subset of attributes

Feature Subset Selection

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: {} => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  A_4? Y N A_1? A_6? Y N Y N Class 1 Class 2 Class 1 Class 2 => Reduced attribute set: $\{A_1, A_4, A_6\}$

What is data exploration?

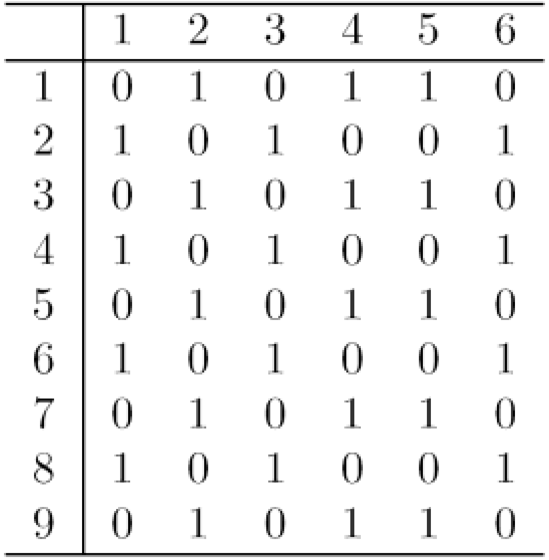
A preliminary exploration of the data to better understand its characteristics.

- Key motivations of data exploration include
  - Humans have a well-developed ability to analyze large amounts of information presented visually
  - Can help detect general patterns and trends
  - Can help detect outliers and unusual patterns

Visualization of data is one of the most powerful and appealing techniques for data exploration.

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

Iris Sample Data Set

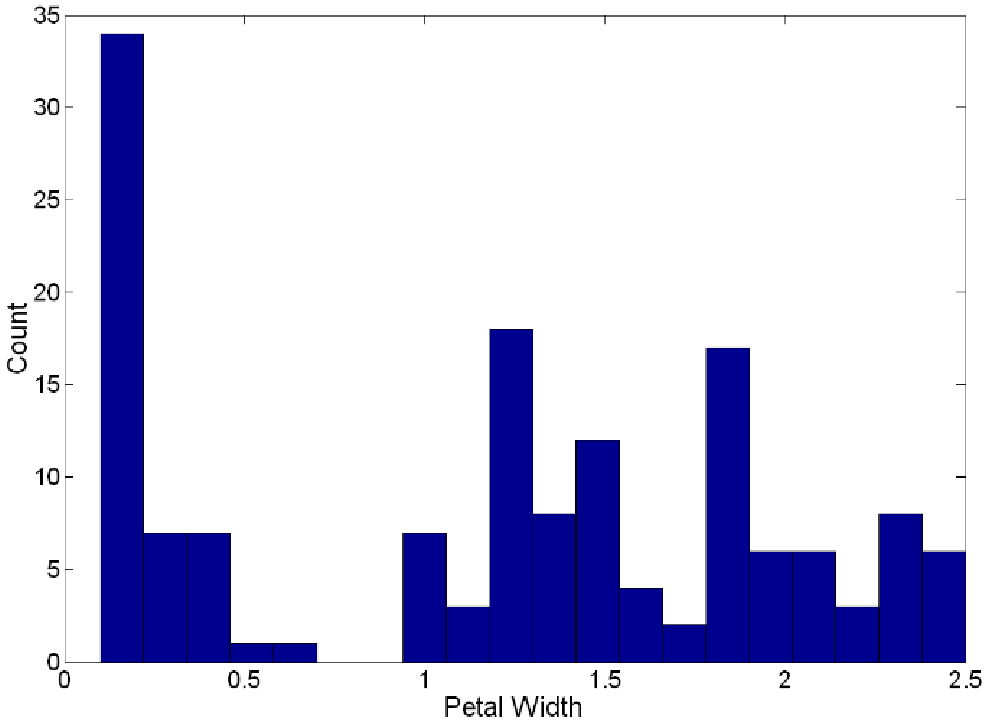
- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html

- 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.

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

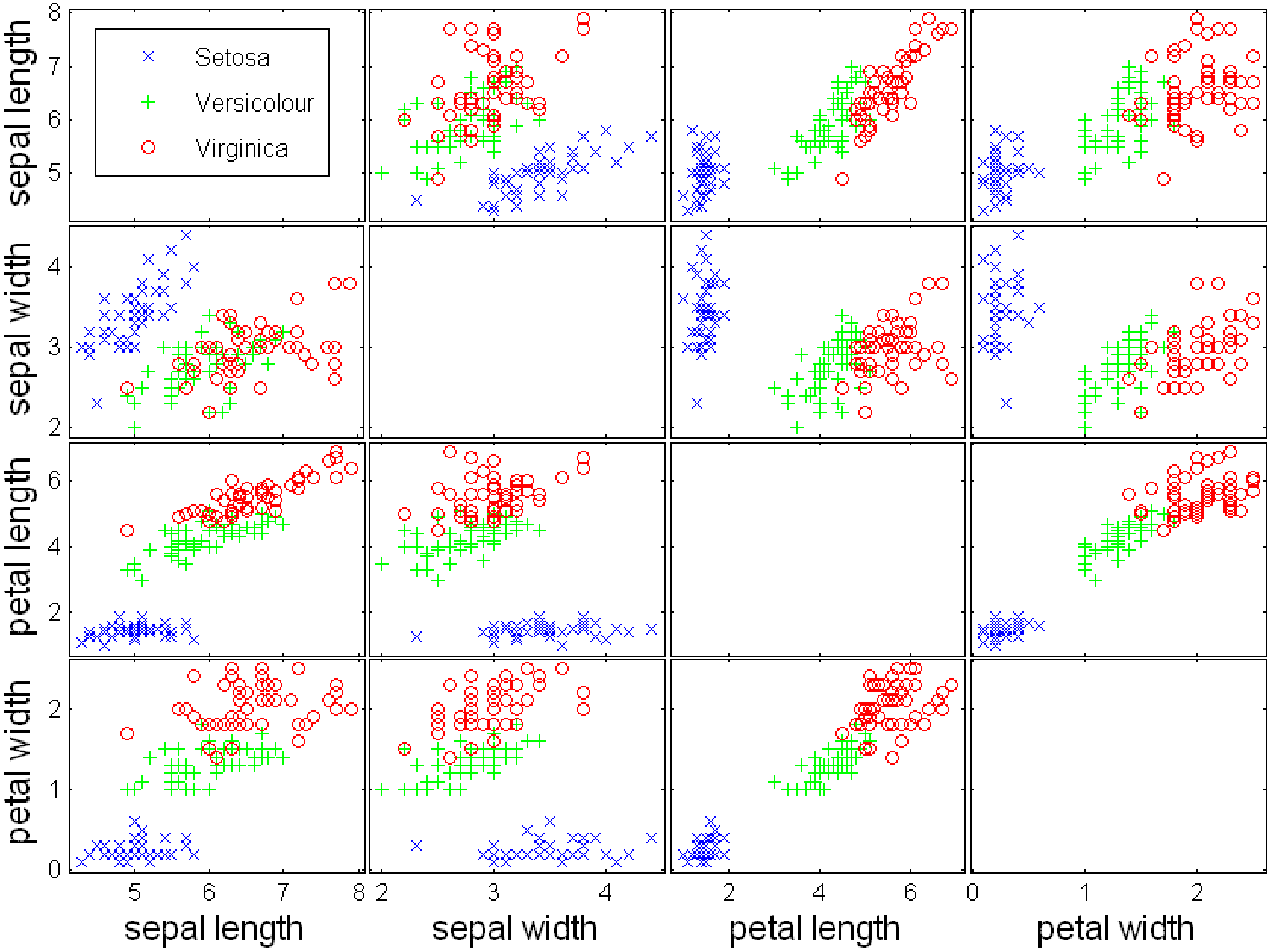


Example: Petal Width

(10 and 20 bins, respectively)

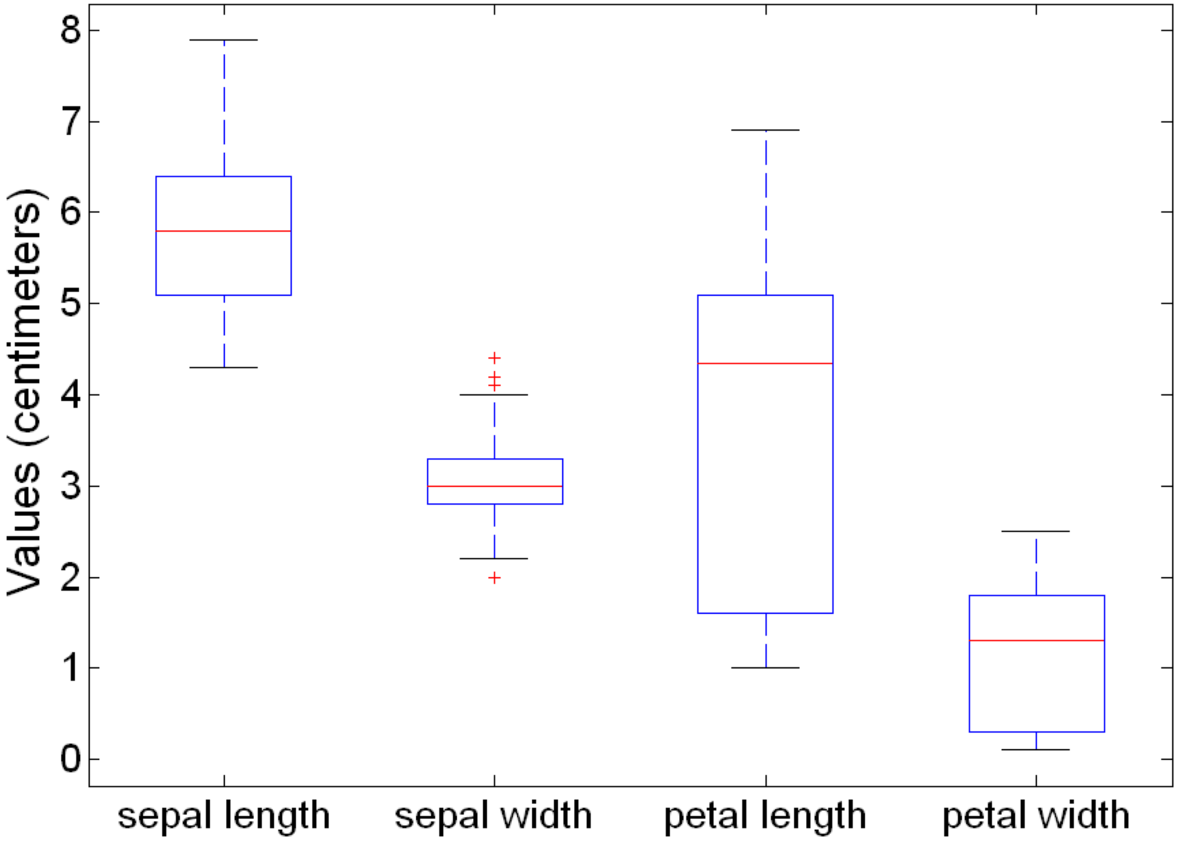
Visualization Techniques: Scatter Plots

Scatter Plot Array of Iris Attributes



Example of Box Plots

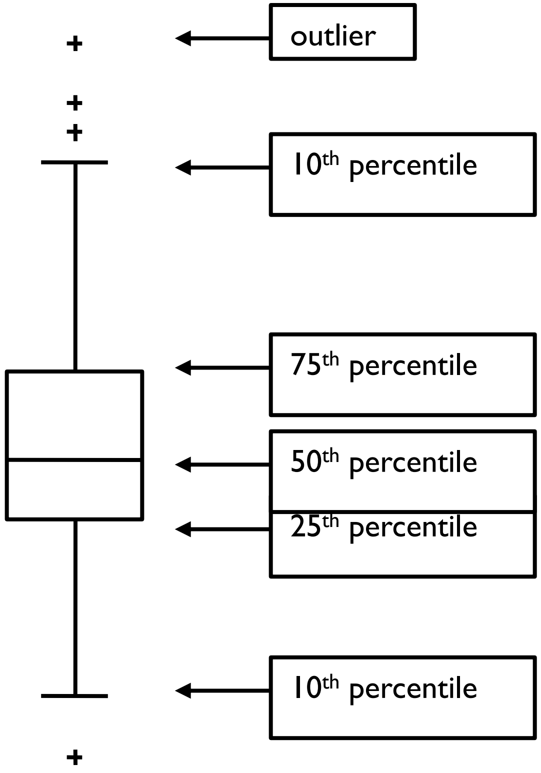
- Box plots can be used to compare attributes



Visualization Techniques: Box Plots

Box Plots

- Boxplots are a popular way of visualizing a distribution.
- Following figure shows the basic part of a box plot



A box plot provides information about an attribute

- range
- median
- normality of the distribution
- skew of the distribution
- plot extreme cases within the sample

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

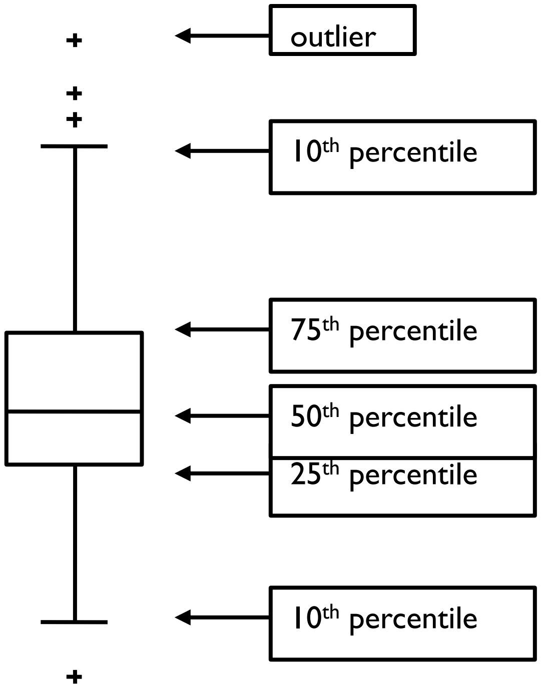
For instance, the 50th percentile is the value such that 50% of all values of x are less than it .

Box Plots Example

- A boxplot incorporates the five-number

  (Minimum,Q1, Median, Q3, Maximum)

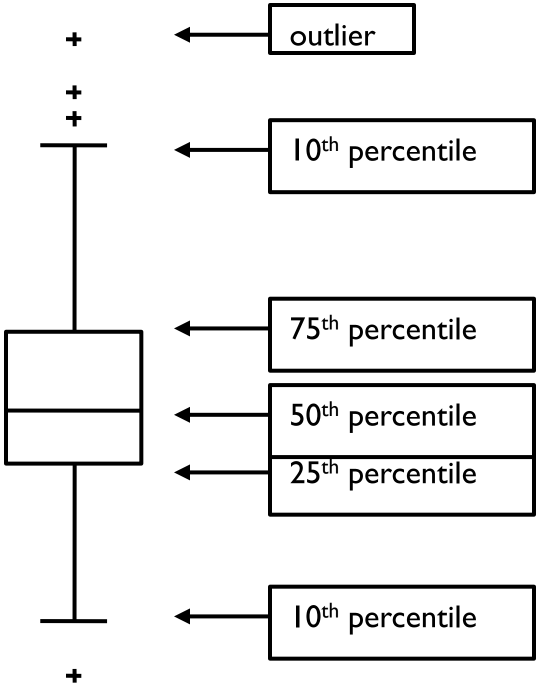
- Typically, the ends of the box are at the quartiles, so that the box length is the interquartile range, *IQR*.
- The **median** is marked by a line within the box
- Two lines (called *whiskers*) outside the box extend to the smallest (*Minimum*) and largest (*Maximum*) observations.



Box Plots Example

When dealing with a moderate number of observations, it is worthwhile to plot **potential outliers individually**.

- To do this in a boxplot, *the whiskers are extended to the extreme low and high observations only if these values are less than* ***1.5×IQR*** *beyond the quartiles.*
- Otherwise, the whiskers terminate at the most extreme observations occurring within 1.5×*IQR* of the quartiles. The remaining cases are plotted individually.



Box Plots Example

Attribute values: 6 47 49 15 42 41 7 39 43 40 36

Sorted: 6 7 15 36 39 40 41 42 43 47 49

Box Plots Example

Attribute values: 6 47 49 15 42 41 7 39 43 40 36

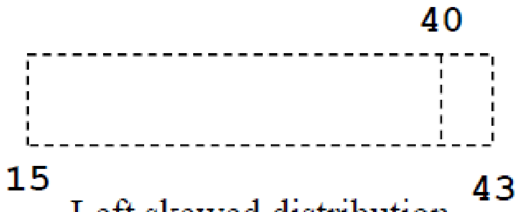
Sorted: 6 7 15 36 39 40 41 42 43 47 49

$Q_1 = 15$ lower quartile

Q_2 = median = 40 *(mean = 33.18)*

$Q_3 = 43$ upper quartile

$Q_3 - Q_1 = 28$ interquartile range



Left skewed distribution

Practice Question

- Introduction to Data Mining
  - Chapters 1 and 2

- Pre-processing using Python (hw/lab/tutorial) will be uploaded by Friday