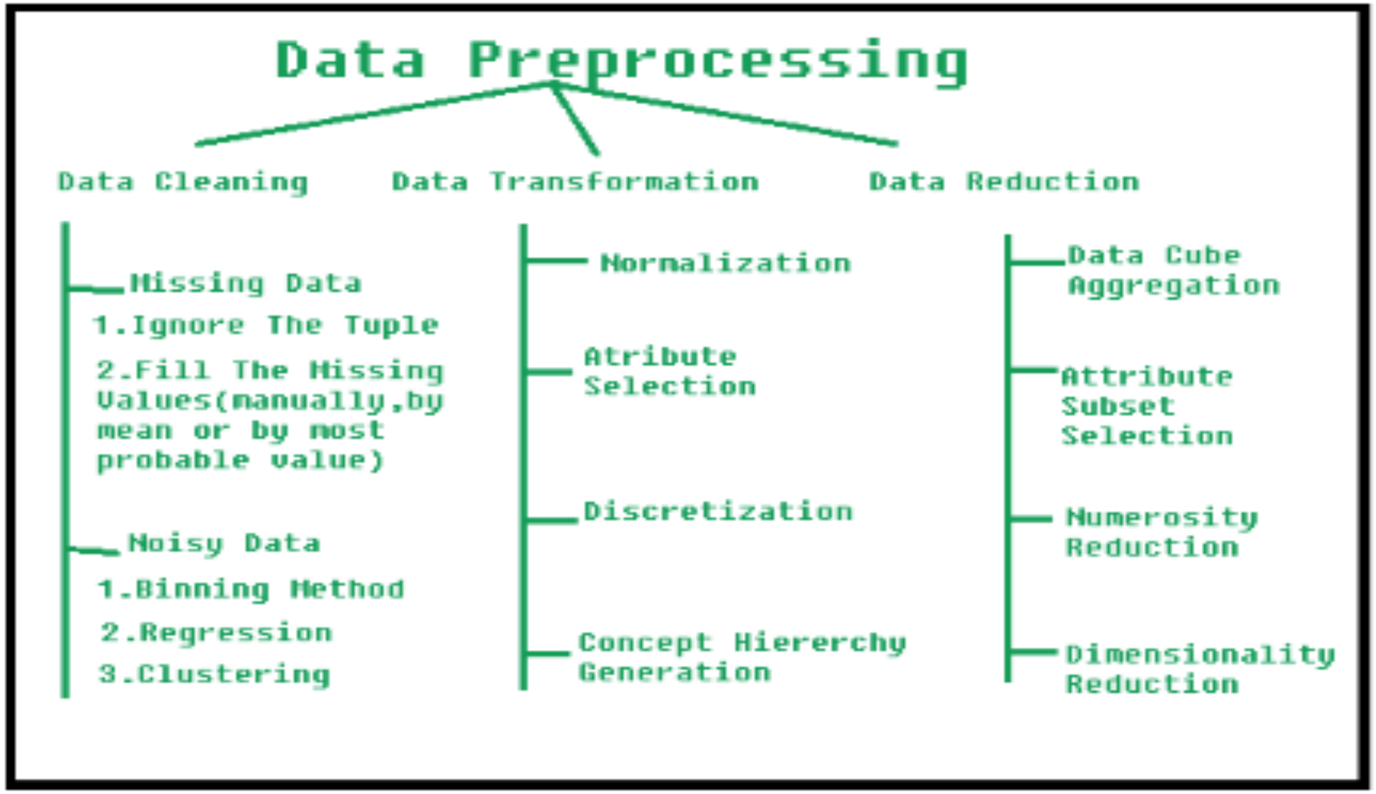
CENG3521 - Data Mining

LECTURE 2

DATA PREPROCESSING

- Data preprocessing is a data mining technique which is used to transform the raw data in a **useful and efficient** format



Data Preprocessing

Data Cleaning

Missing Data

1.Ignore The Tuple

2.Fill The Missing Values(manually,by mean or by most probable value)

Noisy Data

1.Binning Method

2.Regression

3.Clustering

Data Transformation

Normalization

Atribute Selection

Discretization

Concept Hiererchy Generation

Data Reduction

Data Cube Aggregation

Attribute Subset Selection

Numerosity Reduction

Dimensionality Reduction

DATA PREPROCESSING

1. DATA CLEANING

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

a. Missing Data

This situation arises when some data is missing in the data. It can be handled in various ways:

i. Ignore the tuples

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

ii. Fill the missing values

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value

DATA PREPROCESSING

b. Noisy Data

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways:

i. Binning Method

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task.

DATA PREPROCESSING

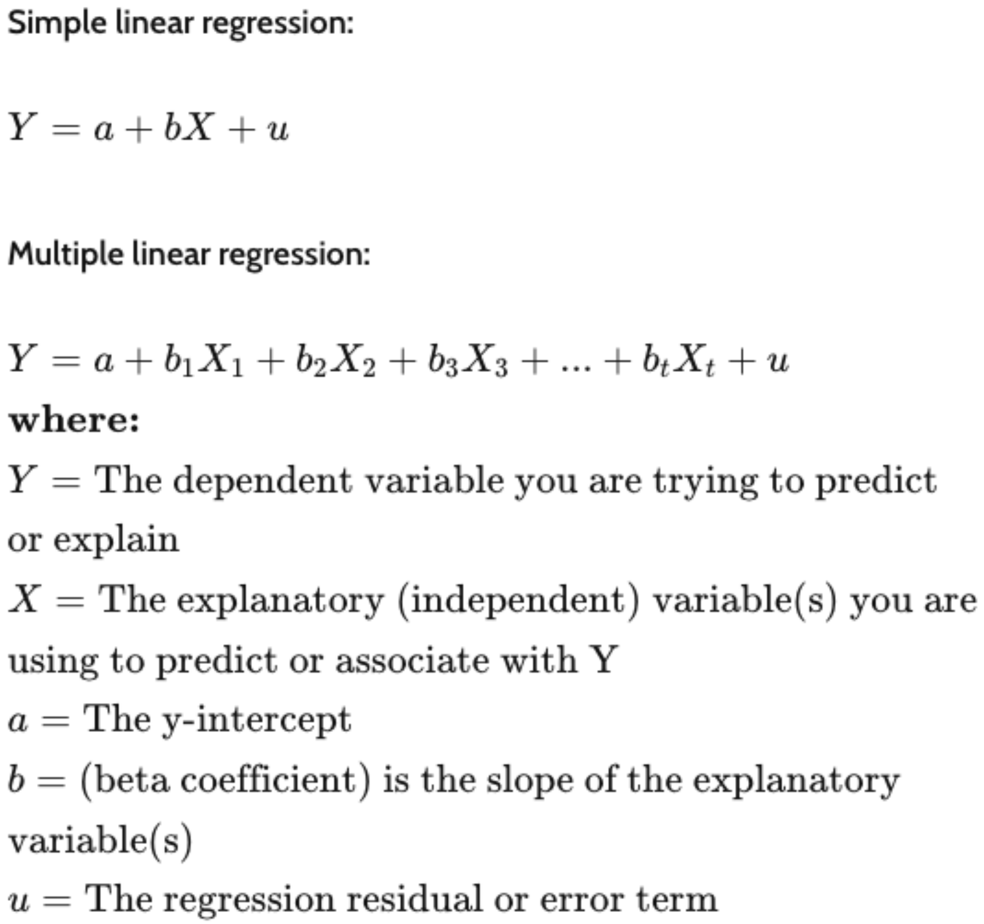
ii. Regression

Here data can be made smooth by fitting it to a regression function. The regression used may be simple linear (having one independent variable) or multiple linear (having multiple independent variables) regression.

Regression captures the correlation between variables observed in a data set and quantifies whether those correlations are statistically significant or not.

DATA PREPROCESSING

ii. Regression



Simple linear regression:

$$Y = a + bX + u$$

Multiple linear regression:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_tX_t + u$$

where:

$Y =$ The dependent variable you are trying to predict or explain

$X =$ The explanatory (independent) variable(s) you are using to predict or associate with Y

$a =$ The y-intercept

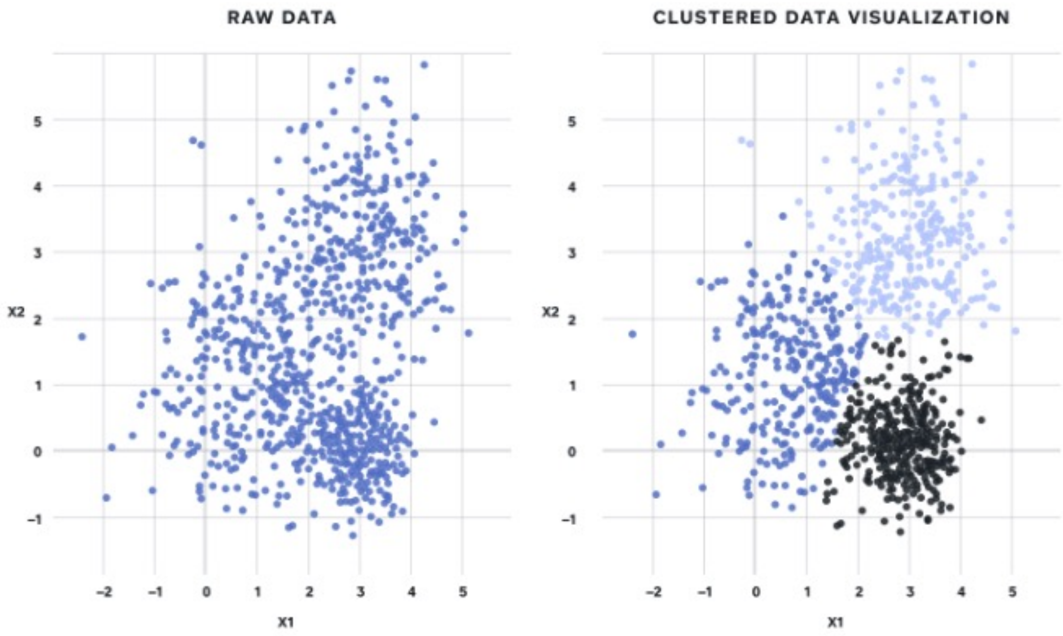
$b =$ (beta coefficient) is the slope of the explanatory variable(s)

$u =$ The regression residual or error term

DATA PREPROCESSING

iii. Clustering

This approach groups the similar data in a cluster. The outliers may be undetected, or it will fall outside the clusters.



RAW DATA

CLUSTERED DATA VISUALIZATION

DATA PREPROCESSING

2. DATA TRANSFORMATION

This step is taken in order to transform the data in appropriate forms suitable for mining process

a. Normalization

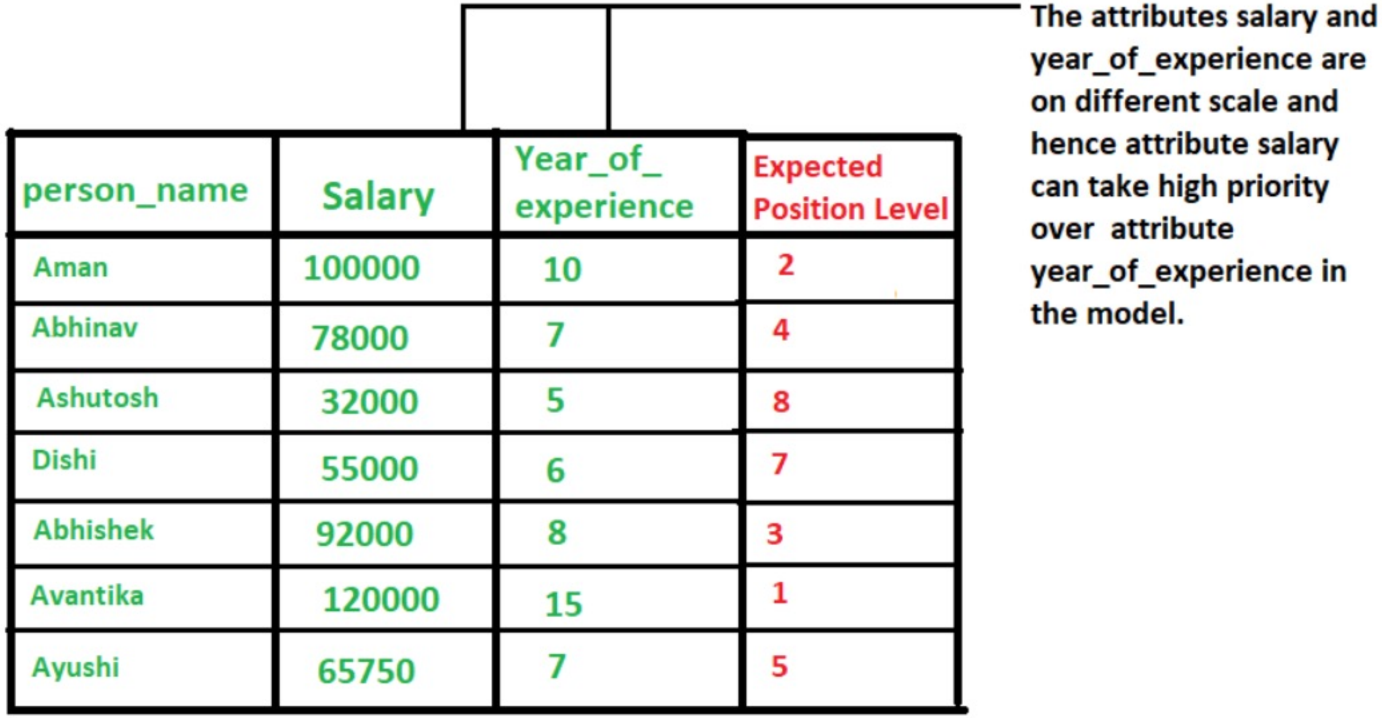
The goal of normalization is to make every datapoint have **the same scale** so ***each feature is equally important***.

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

DATA PREPROCESSING

a. Normalization

The goal of normalization is to make every datapoint have **the same scale** so ***each feature is equally important***.



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

The attributes salary and year_of_experience are on different scale and hence attribute salary can take high priority over attribute year_of_experience in the model.

DATA PREPROCESSING

a. Normalization

i. Min Max

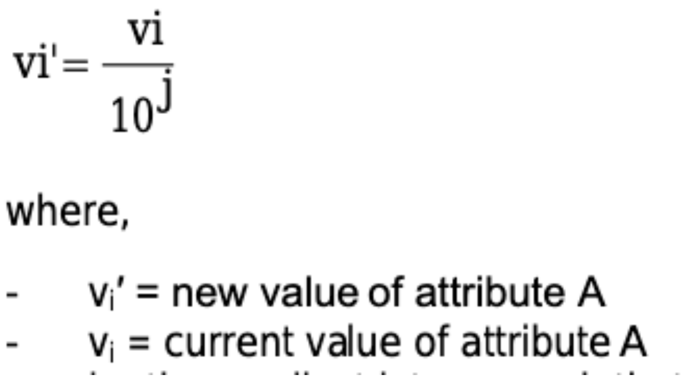
- For every feature, **the minimum value** of that feature **gets transformed into a 0**,
- the **maximum value gets transformed into a 1**, and every other value gets
- transformed into a decimal between 0 and 1.

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

DATA PREPROCESSING

ii. Decimal Scaling

- It functions by converting a number to a decimal point.



$$vi' = \frac{vi}{10^j}$$

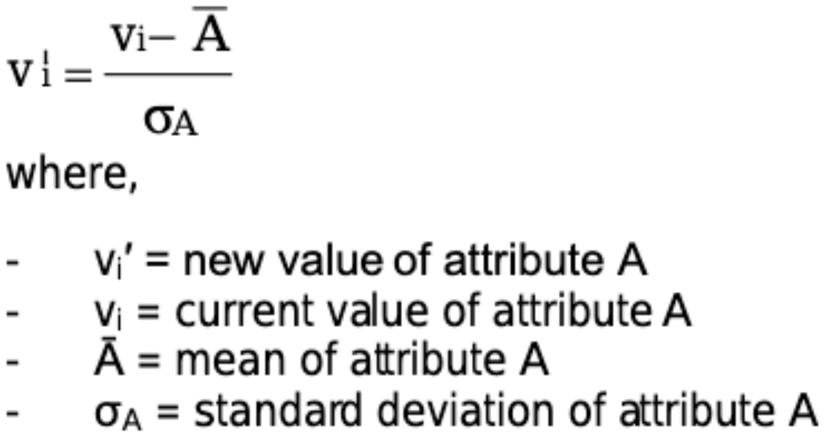
where,

- v_i' = new value of attribute A
- v_i = current value of attribute A

DATA PREPROCESSING

iii. Z score

- Z-score normalization is a strategy of normalizing data that avoids the outlier issue.
- Z-Score value is to understand **how far the data point is from the mean**.
- Technically, it **measures the standard deviations below or above the mean**.



$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

where,

- v_i' = new value of attribute A
- v_i = current value of attribute A
- $\bar{A}$ = mean of attribute A
- σ_A = standard deviation of attribute A

DATA PREPROCESSING

2. DATA TRANSFORMATION

b. Attribute Selection (Attribute subset selection, Feature selection)

- Attribute subset Selection is a technique which is used for **data reduction** in data mining process.
- Data reduction reduces the size of data so that it can be used for analysis purposes **more efficiently**.
- The data set may have a large number of attributes. But some of those attributes can be **irrelevant or redundant**.
  - **Relevant**: These are attributes which have an influence on the output and their role cannot be assumed by the rest.
  - **Irrelevant:** Irrelevant attributes are defined as those attributes not having any influence on the output whose values are generated at random for each example.
  - **Redundant:** A redundant exists, whenever a feature can take the role of another (perhaps the simplest way to model redundancy).

DATA PREPROCESSING

2. DATA TRANSFORMATION

b. Attribute Selection (Attribute subset selection, Feature selection)

- The goal of attribute subset selection is to find a **minimum set of attributes** such that dropping of those irrelevant attributes <u>does not affect the quality of data and the cost of data analysis could be reduced</u>.
- Mining on a reduced data set also makes the **discovered pattern easier to understand.**
- The best way to do the task is to use the **statistical significance tests** such that best (or worst) attributes can be recognized.
- Statistical significance test **assumes that attributes are independent of one another.**

DATA PREPROCESSING

2. DATA TRANSFORMATION

b. **Attribute Selection** (Attribute subset selection, Feature selection)

- This is a kind of greedy approach in which a significance level is decided (statistically ideal value of **significance level is 5%)** and the models are tested again and again until **p-value (probability value) of all attributes is less than or equal** to the selected significance level.
- The attributes having **p-value higher** than significance level are **discarded**.
- This procedure is repeated until **all the attribute in data set has p-value less than or equal to the significance level**.
- This gives us the reduced data set having no irrelevant attributes.

DATA PREPROCESSING

2. DATA TRANSFORMATION

b. Attribute Selection (Attribute subset selection, Feature selection)

Methods of Attribute Subset Selection

1. Stepwise Forward Selection
2. Stepwise Backward Elimination
3. Combination of Forward Selection and Backward Elimination
4. Decision Tree Induction

The "best" (and "worst") attributes are typically determined using: – the tests of **statistical significance**, which assume that the attributes are independent of one another. – the **information gain measure** used in building decision trees for classification.

DATA PREPROCESSING

2. DATA TRANSFORMATION

b. Attribute Selection (Attribute subset selection, Feature selection)

Methods of Attribute Subset Selection

1. Stepwise Forward Selection

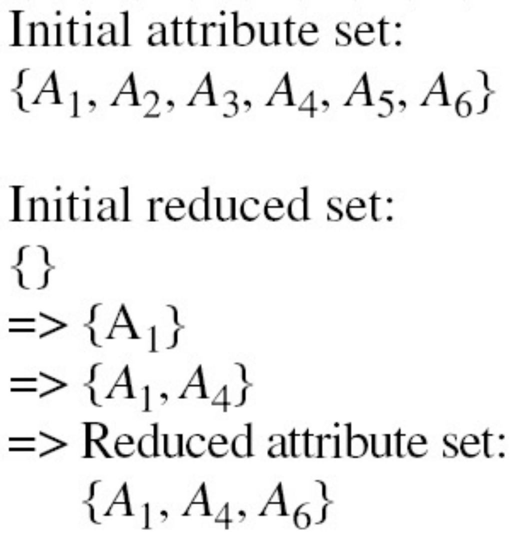
This procedure start with an empty set of attributes as the minimal set. The most relevant attributes are chosen(having minimum p-value) and are added to the minimal set. In each iteration, one attribute is added to a reduced set.

DATA PREPROCESSING

2. DATA TRANSFORMATION

1. Stepwise Forward Selection

- The procedure starts with an empty set of attributes as the reduced set.
- First: The best single-feature is picked.
- Next: At each subsequent iteration or step, the best of the remaining original attributes is added to the set.



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:
$\{A_1, A_4, A_6\}$

DATA PREPROCESSING

2. DATA TRANSFORMATION

b. Attribute Selection (Attribute subset selection, Feature selection)

Methods of Attribute Subset Selection

2. Stepwise Backward Elimination

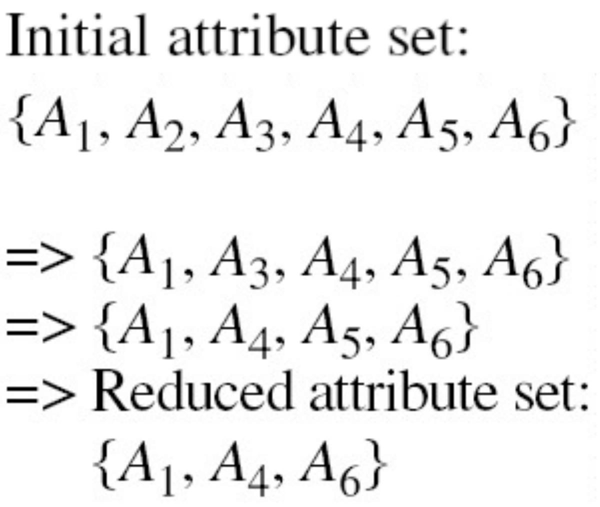
Here all the attributes are considered in the initial set of attributes. In each iteration, one attribute is eliminated from the set of attributes whose p-value is higher than significance level.

DATA PREPROCESSING

2. DATA TRANSFORMATION

2. Stepwise Backward Elimination

- The procedure starts with the full set of attributes.
- At each step, it removes the worst attribute remaining in the set.



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:
$\{A_1, A_4, A_6\}$

DATA PREPROCESSING

2. DATA TRANSFORMATION

b. Attribute Selection (Attribute subset selection, Feature selection)

Methods of Attribute Subset Selection

3. Combination of Forward Selection and Backward Elimination

The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently. This is the most common technique which is generally used for attribute selection.

DATA PREPROCESSING

2. DATA TRANSFORMATION

b. **Attribute Selection** (Attribute subset selection, Feature selection)

Methods of Attribute Subset Selection

4. Decision Tree Induction

- Decision tree algorithms, such as ID3, C4.5, and CART, were originally intended for classification.
- Decision tree induction constructs a flow chart-like structure where each internal (non-leaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction.

DATA PREPROCESSING

2. DATA TRANSFORMATION

4. Decision Tree Induction

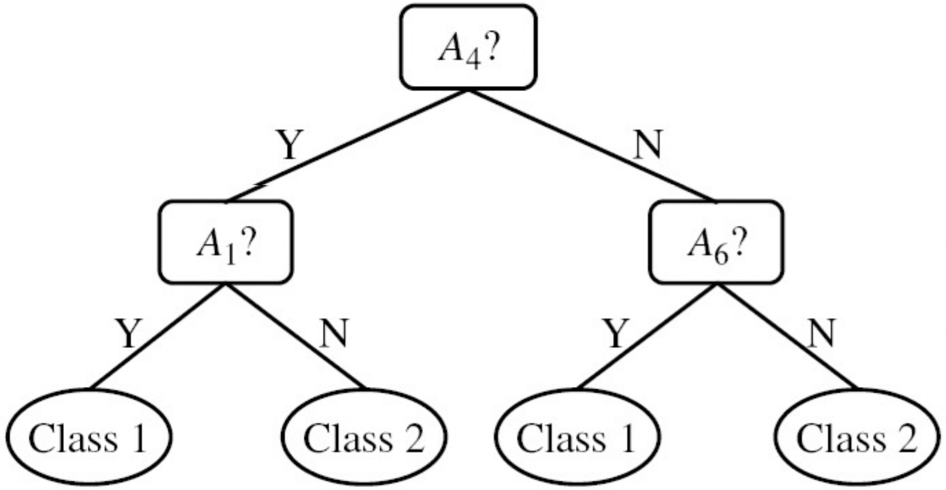
- At each node, the algorithm chooses the "best" attribute to partition the data into individual classes.
- When decision tree induction is used for attribute subset selection, a tree is constructed from the given data.
- All attributes that do not appear in the tree are assumed to be irrelevant.

DATA PREPROCESSING

2. DATA TRANSFORMATION

4. Decision Tree Induction

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$



=> Reduced attribute set:
$\{A_1, A_4, A_6\}$

DATA PREPROCESSING

2. DATA TRANSFORMATION

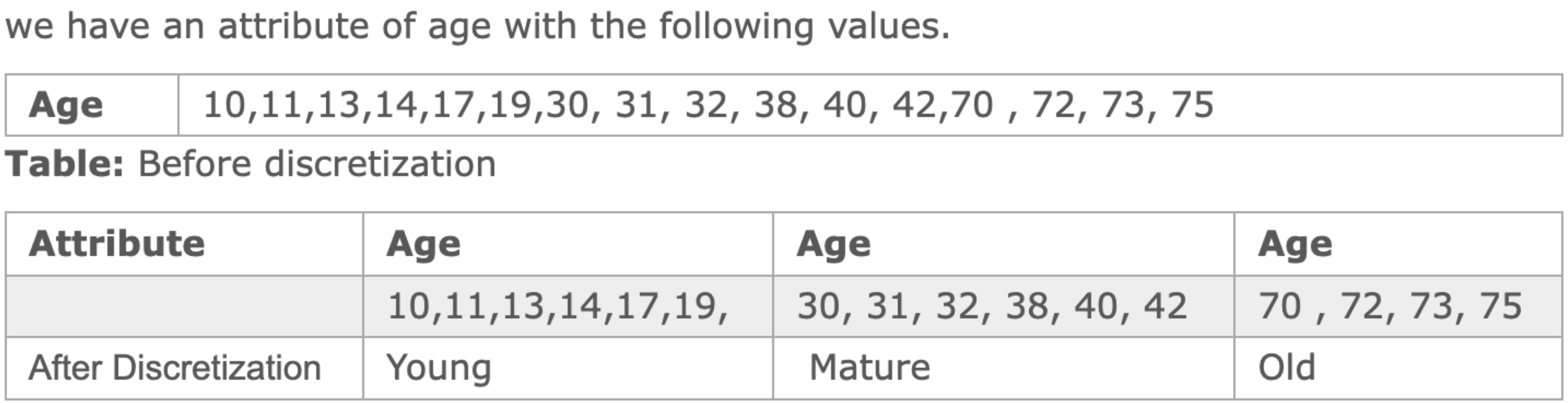
c. Discretization

- This is a process of converting **continuous data** into a set of data **intervals**.
- Continuous attribute values are substituted by small interval labels.
- This makes the data easier to study and analyze.
- If a continuous attribute is handled by a data mining task, then its discrete values can be replaced by constant quality attributes. This **improves the efficiency** of the task.

DATA PREPROCESSING

2. DATA TRANSFORMATION

c. Discretization



we have an attribute of age with the following values.

Age	10,11,13,14,17,19,30, 31, 32, 38, 40, 42,70 , 72, 73, 75

Table: Before discretization

Attribute	Age	Age	Age
	10,11,13,14,17,19,	30, 31, 32, 38, 40, 42	70 , 72, 73, 75
After Discretization	Young	Mature	Old

DATA PREPROCESSING

2. DATA TRANSFORMATION

c. Discretization

- Unsupervised:
  - Equal-Width Discretization (Binning)
  - Equal-Depth (Frequency) Discretization (Binning)
  - K-Means
- Supervised:
  - Decision Trees

DATA PREPROCESSING

- **Equal-width (distance) Discretization**

- Divides the range into N intervals of equal size: uniform grid

- if A and B are the lowest and highest values of the attribute, the width of

- intervals will be: **W = (B –A)/N.**

- The most straightforward, but outliers may dominate presentation

- Skewed data is not handled well

DATA PREPROCESSING

Equal-width (distance) Discretization

Example:

Sorted data for price (in dollars):

4, 8, 15, 21, 21, 24, 25, 28, 34

W = (B –A)/N = (34 – 4) / 3 =10

Bin 1: 4-14,
Bin2: 15-24,
Bin 3: 25-34

Bin 1: 4, 8
Bin 2: 15, 21, 21, 24
Bin 3: 25, 28, 34

DATA PREPROCESSING

- **Equal-depth (frequency) Discretization**

- Divides the range into N intervals, each containing approximately same number of samples

- Good data scaling

DATA PREPROCESSING

Equal-depth (frequency) discretization

Example:

Sorted data for price (in dollars):

4, 8, 15, 21, 21, 24, 25, 28, 34

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

DATA PREPROCESSING

K-means discretization

We apply K-Means clustering to the continuous variable, thus dividing it into discrete groups or clusters.

- K-Means doesn’t improve the value spread
- It can handle outliers; however a centroid bias may exist.
- Can be combined with categorical encoding

DATA PREPROCESSING

2. DATA TRANSFORMATION

c. Discretization

After discretizing variables, you can do either of the following:

- Build decision tree algorithms and directly use the output of discretization as the number of bins. The decision trees can find non-linear relationships between the discretized variable and the target variables.
- Use a linear model, while the bins do not have a linear relationship with the target variable. Improve the model by treating bins as categories with some sort of encoding.

DATA PREPROCESSING

2. DATA TRANSFORMATION

d. Concept Hierarchy Generation

- Discretization can be performed rapidly on an attribute to provide a hierarchical partitioning of the attribute values, known as a ***Concept Hierarchy.***
- Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.

i. CHG for numerical variables

ii. CHG for categorical variables

DATA PREPROCESSING

i. CHG for numerical variables

Typical Methods of Discretization and Concept Hierarchy Generation for Numerical Data:

- Binning
- Histogram analysis
- Cluster analysis
- Entropy-Based Discretization
- Interval Merge by χ2 Analysis

DATA PREPROCESSING

ii. CHG for categorical variables

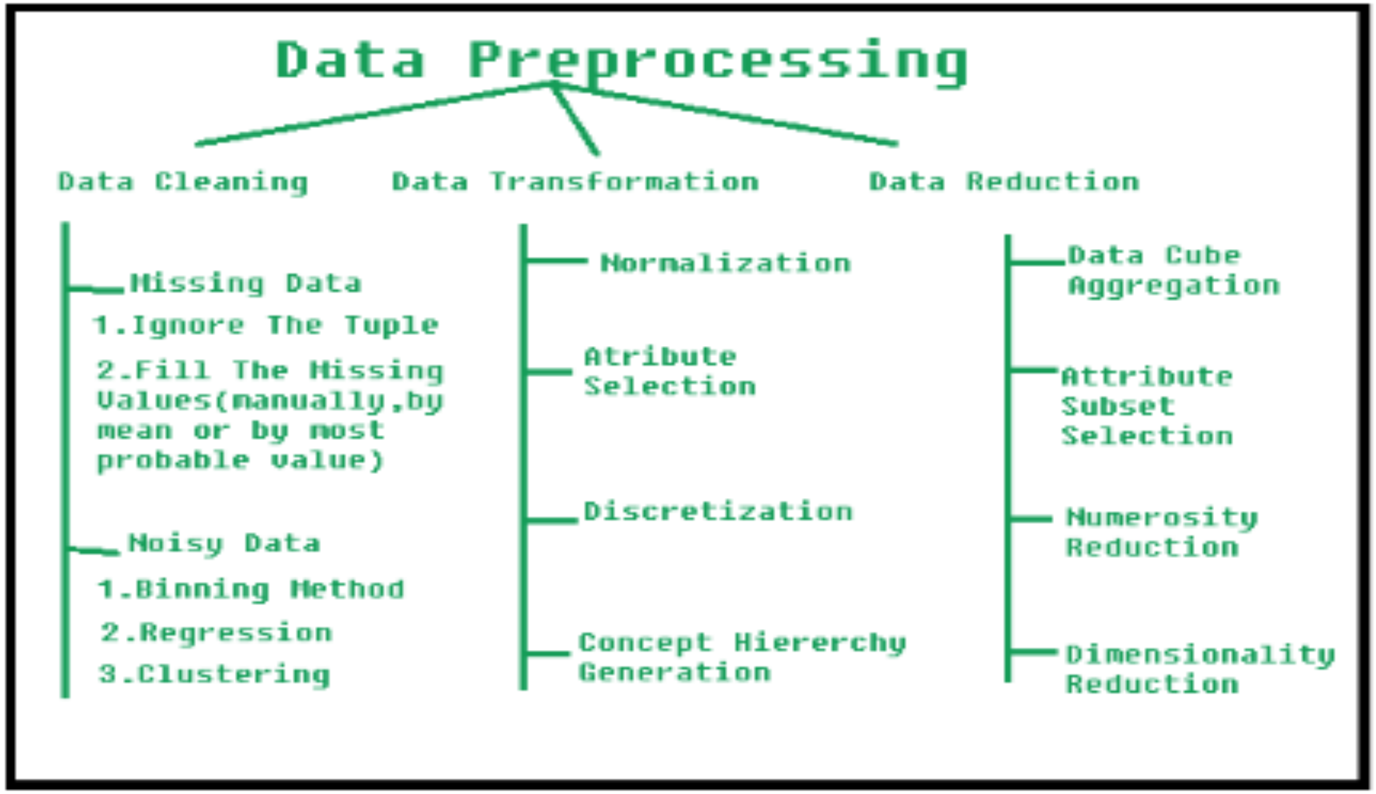
Generalization is the generation of concept hierarchies for categorical data.

- Categorical attributes have a finite (but possibly large) number of distinct values, with no ordering among the values.
- Data Discretization and Concept Hierarchy Generation Examples include:
  - geographic location,
  - job category, and
  - Item type.

i.e. {Urbana, Champaign, Chicago} < Illinois

DATA PREPROCESSING

- Data preprocessing is a data mining technique which is used to transform the raw data in a **useful and efficient** format



Data Preprocessing

Data Cleaning

- Missing Data
  1. Ignore The Tuple
  2. Fill The Missing Values(manually,by mean or by most probable value)
- Noisy Data
  1. Binning Method
  2. Regression
  3. Clustering

Data Transformation

- Normalization
- Atribute Selection
- Discretization
- Concept Hiererchy Generation

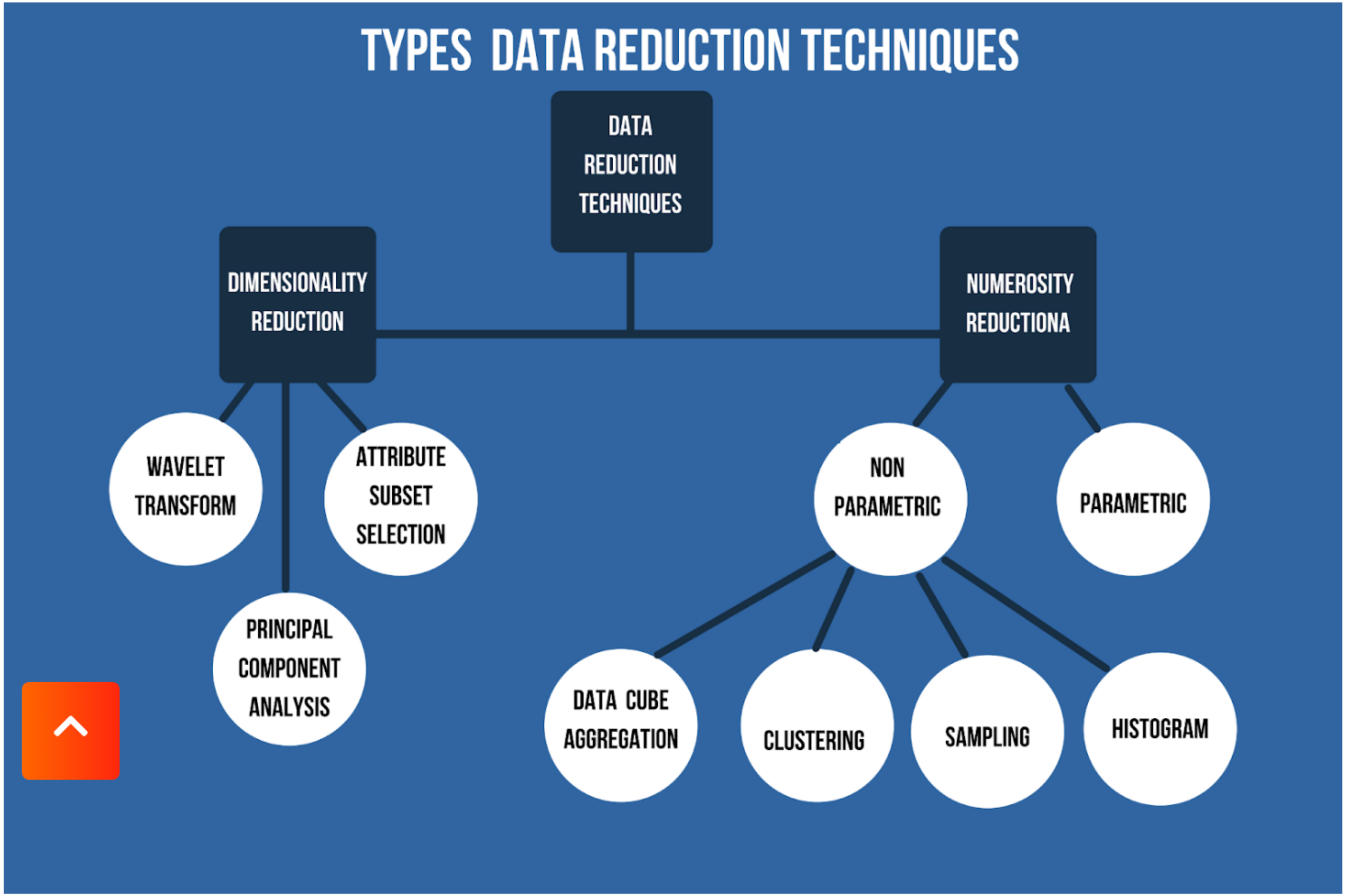
Data Reduction

- Data Cube Aggregation
- Attribute Subset Selection
- Numerosity Reduction
- Dimensionality Reduction

DATA PREPROCESSING

3. DATA REDUCTION

- **Data reduction** is a process that reduced the volume of original data and represents it in a much smaller volume. Data reduction techniques ensure the integrity of data while reducing the data.
- Data reduction does not affect the result obtained from data mining that means the result obtained from data mining before data reduction and after data reduction is the same (or almost the same).



TYPES DATA REDUCTION TECHNIQUES

DATA REDUCTION TECHNIQUES

DIMENSIONALITY REDUCTION

NUMEROSITY REDUCTIONA

WAVELET TRANSFORM

ATTRIBUTE SUBSET SELECTION

PRINCIPAL COMPONENT ANALYSIS

NON PARAMETRIC

PARAMETRIC

DATA CUBE AGGREGATION

CLUSTERING

SAMPLING

HISTOGRAM

DATA PREPROCESSING

3. DATA REDUCTION

a. Dimensionality Reduction

Dimensionality reduction **eliminates the attributes** from the data set under consideration thereby reducing the volume of original data.

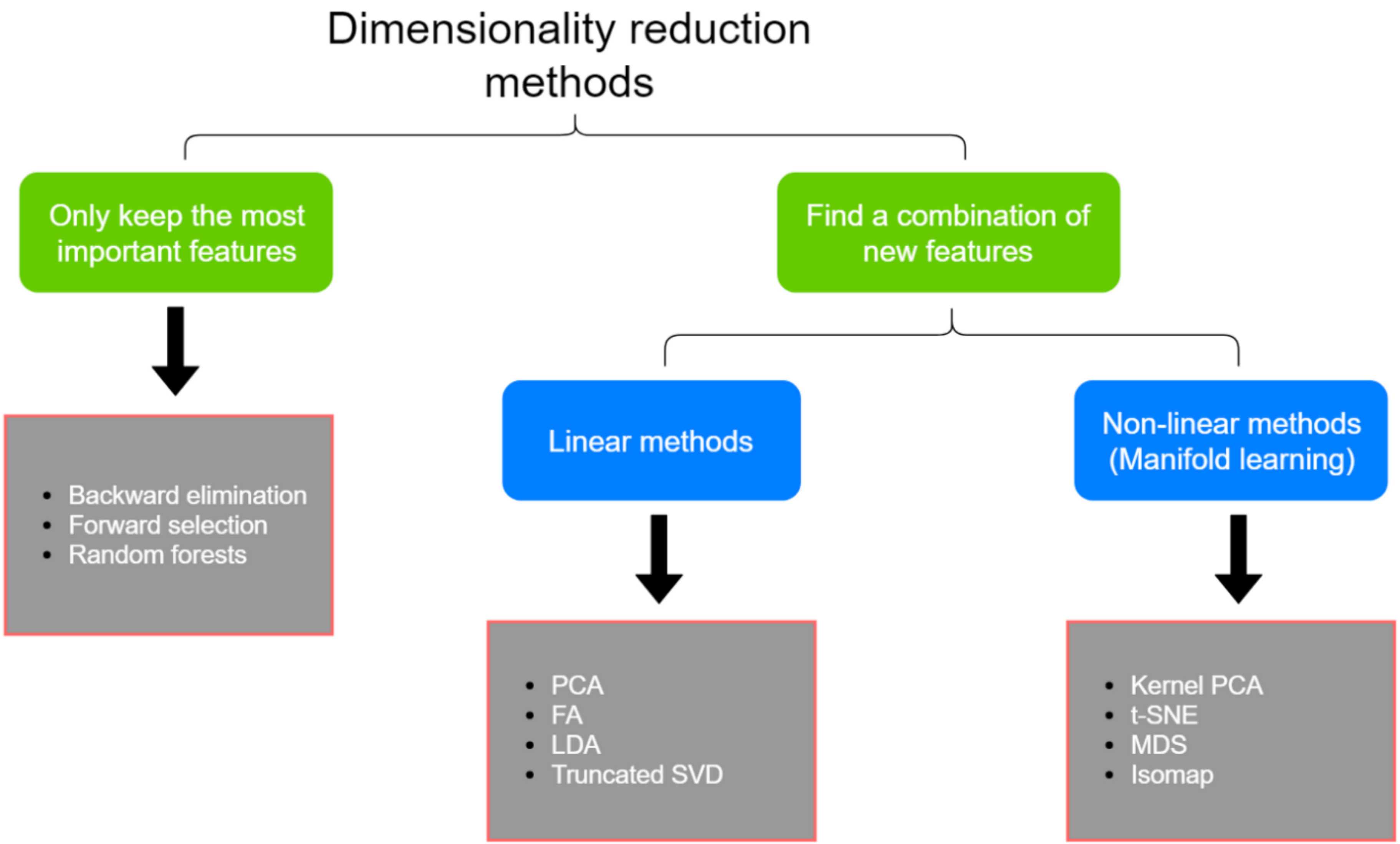
Advantages of Dimensionality Reduction

- A lower number of dimensions in data means less training time and less computational resources and increases the overall performance of machine learning algorithms
- Dimensionality reduction avoids the problem of *overfitting*

DATA PREPROCESSING

Advantages of Dimensionality Reduction

- Dimensionality reduction is extremely useful for *data visualization*
- Dimensionality reduction takes care of *multicollinearity*
- Dimensionality reduction is very useful for *factor analysis*
- Dimensionality reduction removes noise in the data
- Dimensionality reduction can be used to transform non-linear data into a linearly-separable form



Dimensionality reduction methods

Only keep the most important features

- Backward elimination
- Forward selection
- Random forests

Find a combination of new features

Linear methods

- PCA
- FA
- LDA
- Truncated SVD

Non-linear methods (Manifold learning)

- Kernel PCA
- t-SNE
- MDS
- Isomap

DATA PREPROCESSING

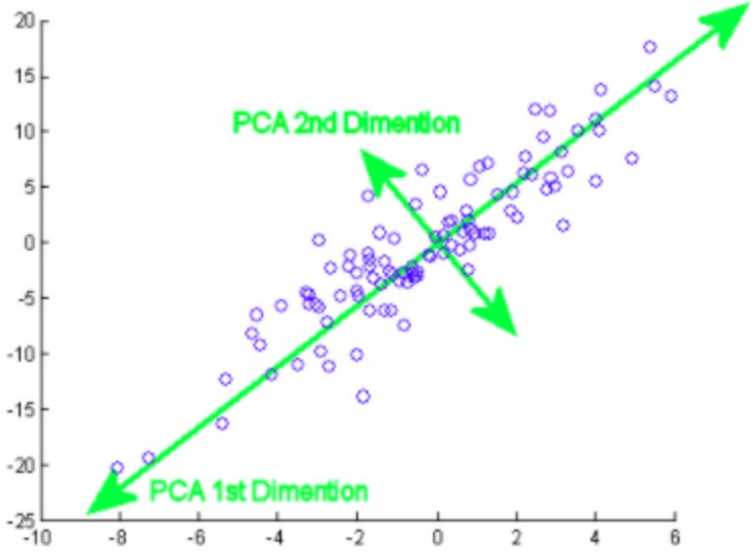
Principal Component Analysis (PCA)

- This method transforms the data into a new coordinate, where the one with the highest variance is the primary principal component. Thus providing us the best possible representations of data.
- It has several advantages, which include reduction of data size(hence faster execution), better visualizations with fewer dimensions, maximizes variance, reduces overfitting, etc.

DATA PREPROCESSING

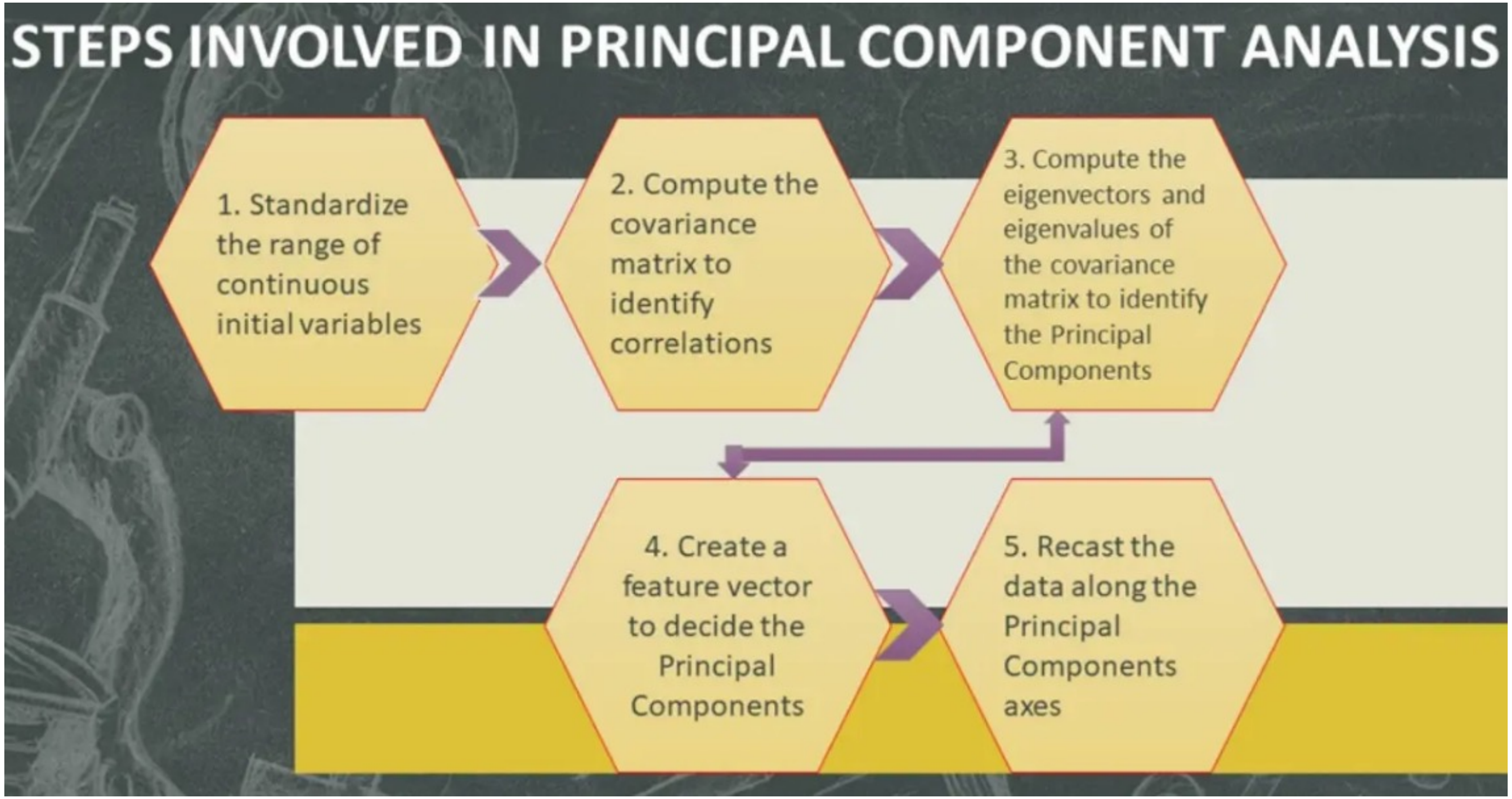
Principal Component Analysis (PCA)

- The principal component means the sequences of direction vectors that differ on basis of best-fit lines. It can also be stated that these components are eigenvectors of the covariance matrix.



DATA PREPROCESSING

Principal Component Analysis (PCA)



STEPS INVOLVED IN PRINCIPAL COMPONENT ANALYSIS

1. Standardize the range of continuous initial variables

2. Compute the covariance matrix to identify correlations

3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the Principal Components

4. Create a feature vector to decide the Principal Components

5. Recast the data along the Principal Components axes

DATA PREPROCESSING

Factor Analysis (FA)

- Factor analysis is the study of unobserved variables, also known as latent variables or latent factors, that may combine with observed variables to affect outcomes

   i. **Exploratory Factor Analysis**

   Exploratory Factor Analysis is used to find the underlying structure of a large set of variables. It reduces data to a much smaller set of summary variables.

   ii. **Confirmatory Factor Analysis**

   Confirmatory Factor Analysis allows you to figure out if a relationship between a set of observed variables and their underlying constructs exists.

DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

- In the Numerosity reduction, the data volume is reduced by choosing an alternative, smaller form of data representation.
- These techniques may be parametric or nonparametric.
- For parametric methods, a model is used to estimate the data, so that only the data parameters need to be stored, instead of the actual data, for example, Log-linear models.
- Non-parametric methods are used for storing a reduced representation of the data which include histograms, clustering, data cube aggregation and sampling.

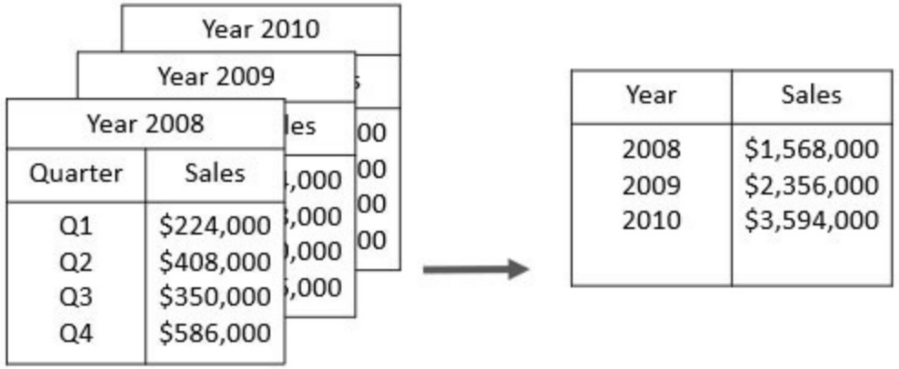
DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

i. Data Cube Aggregation

The data cube aggregation is a multidimensional aggregation which eases multidimensional analysis.


Year 2010
Year 2009

Year 2008

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2008	$1,568,000
2009	$2,356,000
2010	$3,594,000

Aggregated Data

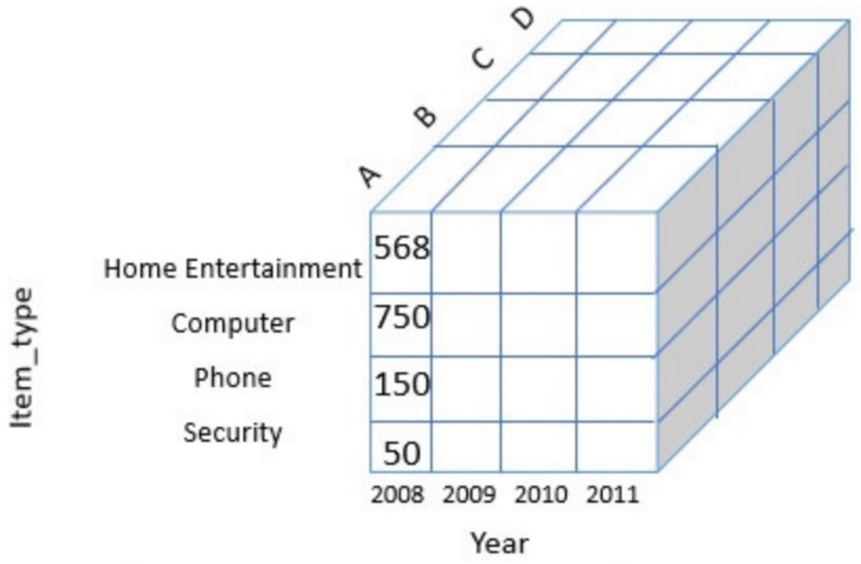
DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

i. Data Cube Aggregation

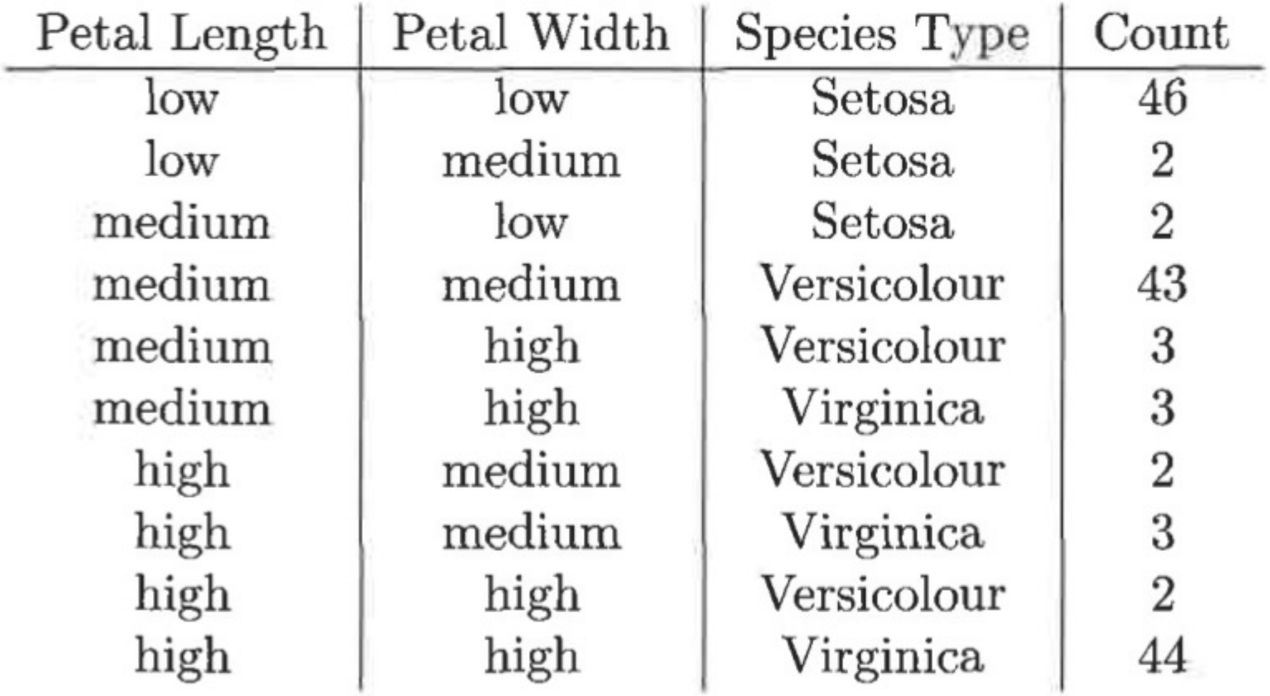
The data cube present precomputed and sum access.



Data Cube Aggregation

DATA PREPROCESSING

Data Cube Aggregation



Petal Length	Petal Width	Species Type	Count
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

DATA PREPROCESSING

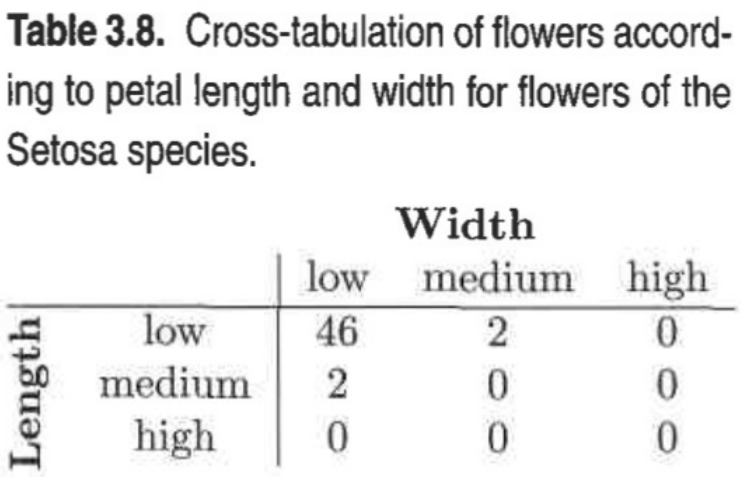


Table 3.8. Cross-tabulation of flowers according to petal length and width for flowers of the Setosa species.

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

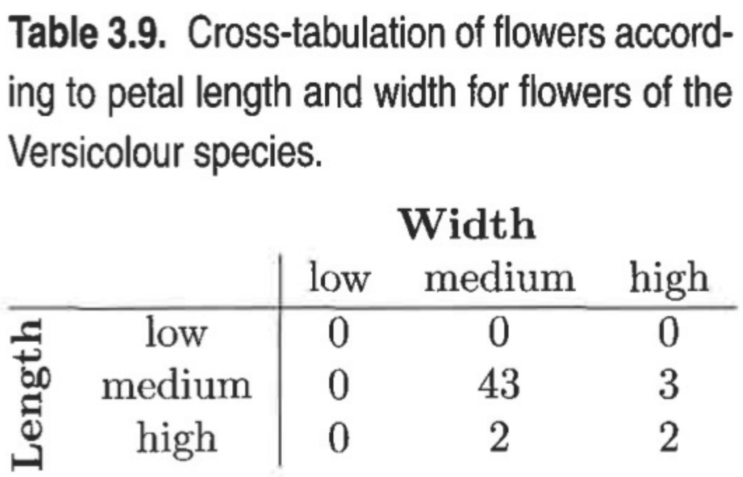


Table 3.9. Cross-tabulation of flowers according to petal length and width for flowers of the Versicolour species.

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

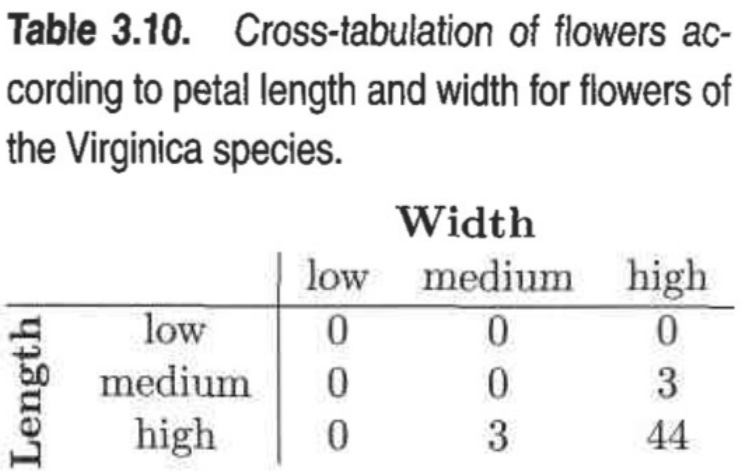
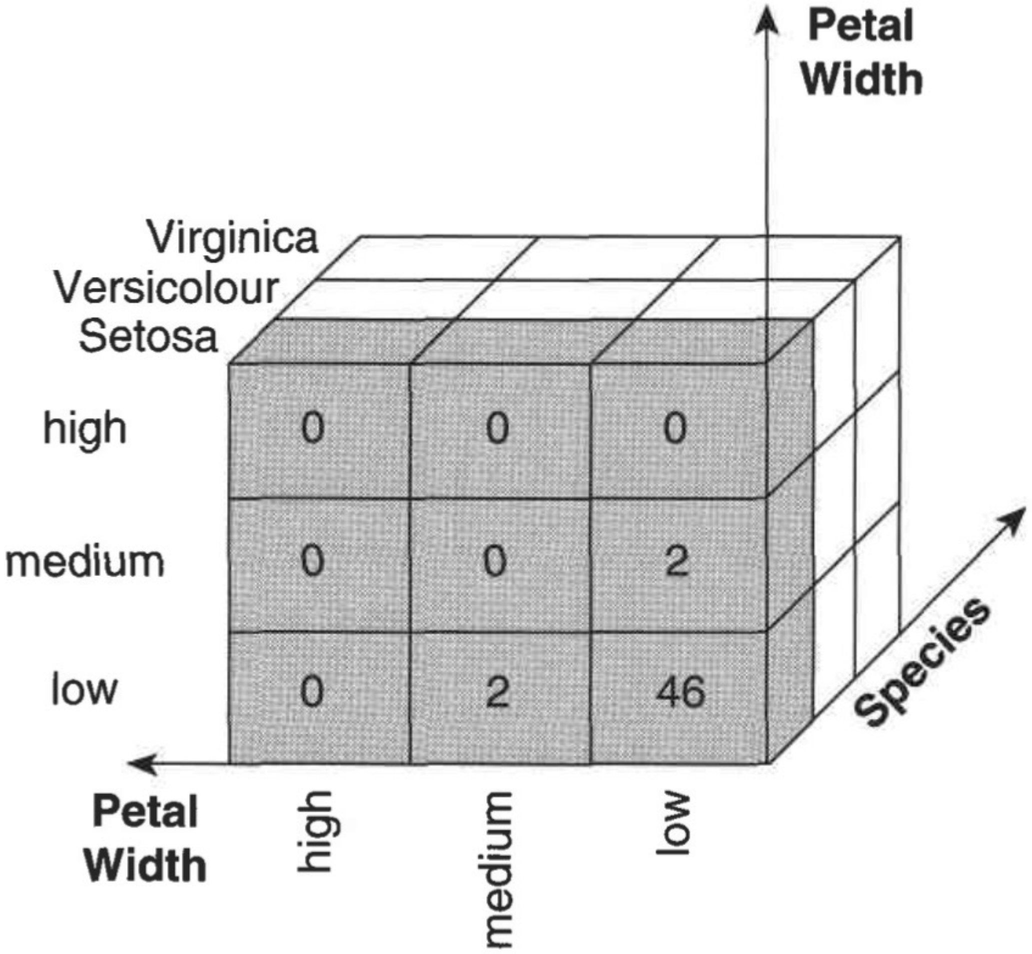


Table 3.10. *Cross-tabulation of flowers according to petal length and width for flowers of the Virginica species.*

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

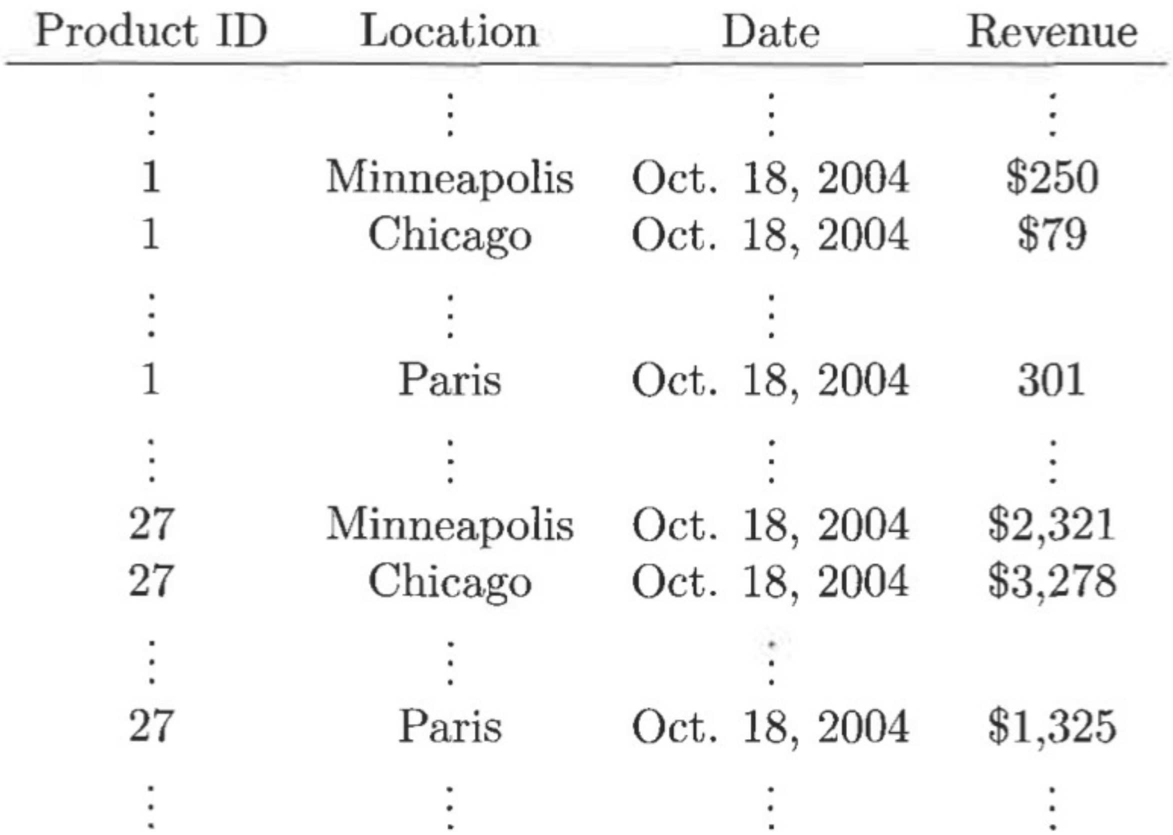
DATA PREPROCESSING



	high	medium	low
high	0	0	0
medium	0	0	2
low	0	2	46

DATA PREPROCESSING

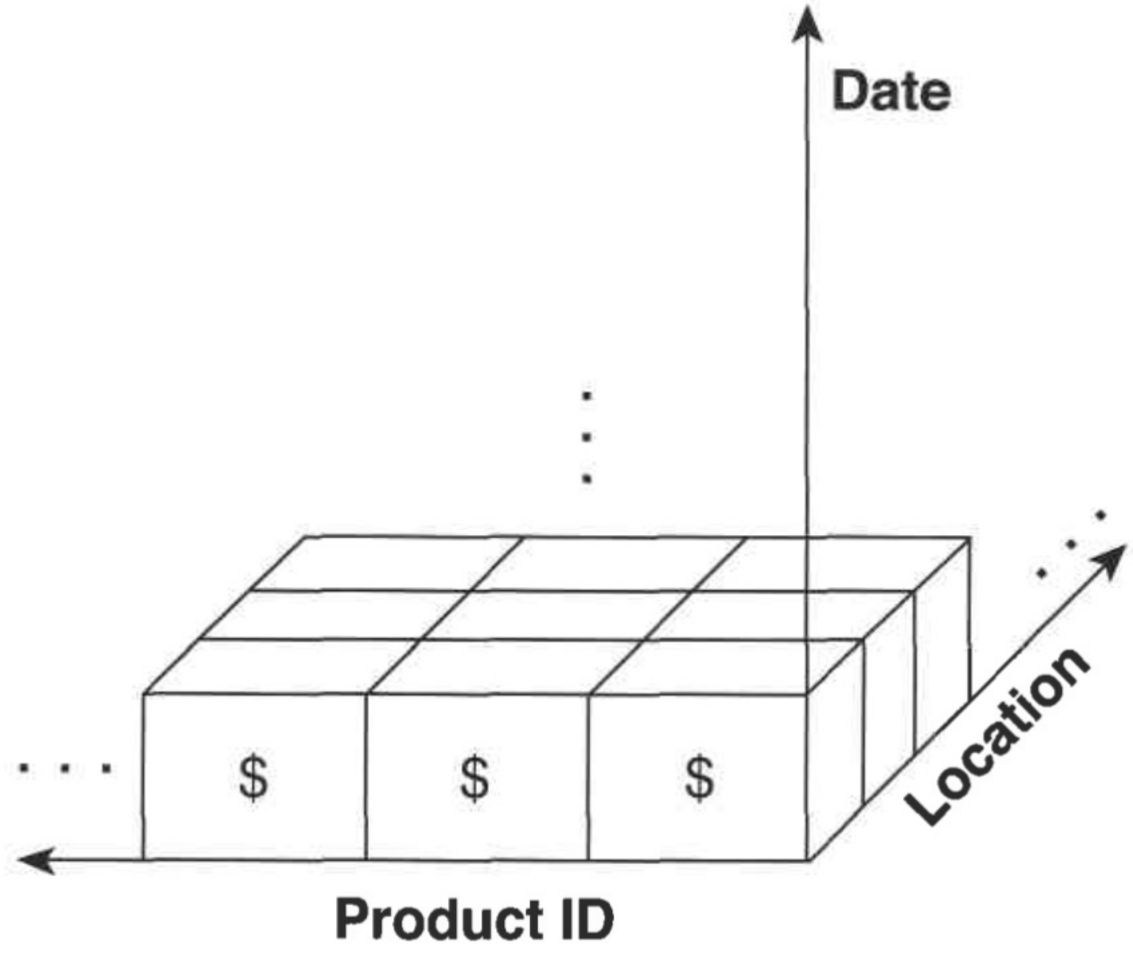
Example



Product ID	Location	Date	Revenue
⋮	⋮	⋮	⋮
1	Minneapolis	Oct. 18, 2004	$250
1	Chicago	Oct. 18, 2004	$79
⋮	⋮	⋮	
1	Paris	Oct. 18, 2004	301
⋮	⋮	⋮	⋮
27	Minneapolis	Oct. 18, 2004	$2,321
27	Chicago	Oct. 18, 2004	$3,278
⋮	⋮	⋮	
27	Paris	Oct. 18, 2004	$1,325
⋮	⋮	⋮	⋮

DATA PREPROCESSING

Example



DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

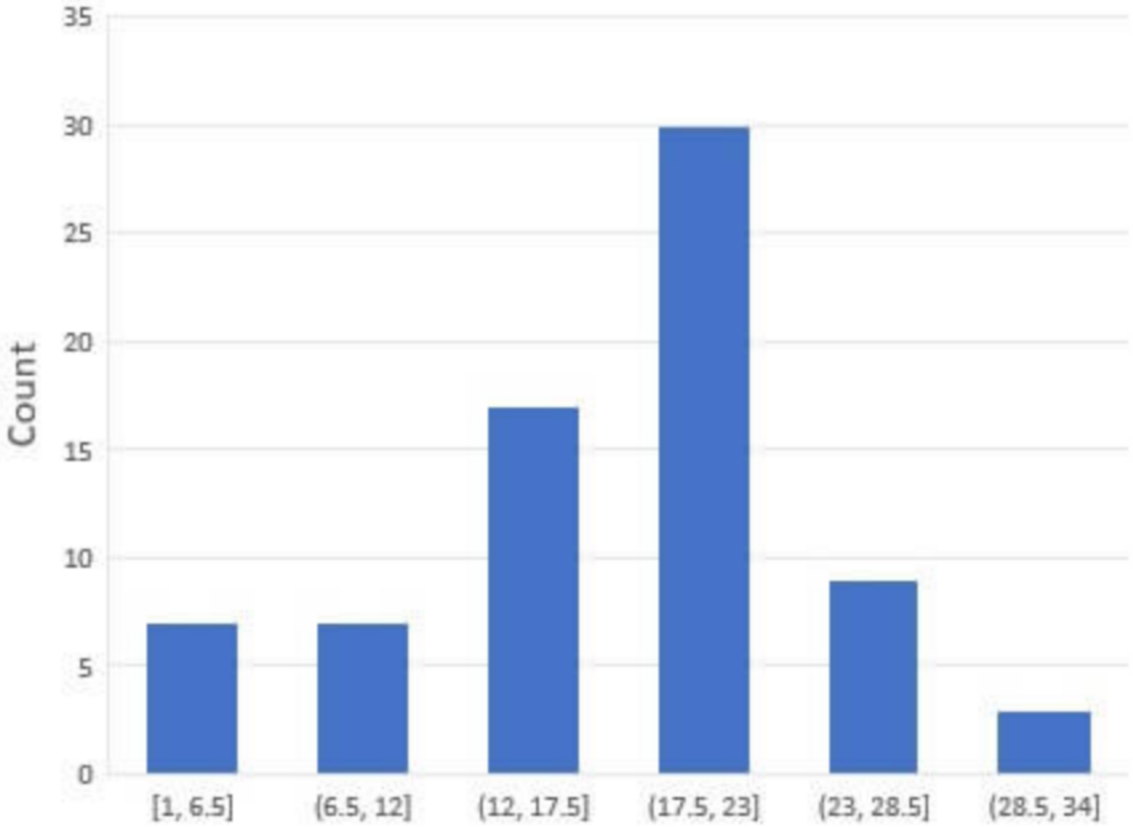
ii. Histogram

- A histogram is a ‘graph’ that represents frequency distribution which describes how often a value appears in the data.
- Histogram uses the binning method and to represent data distribution of an attribute. It uses disjoint subset which we call as bin or buckets.

We have data for *AllElectronics data set,* which contains prices for regularly sold items.

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 34

The diagram below shows a histogram of equal width that shows the frequency of price distribution.



DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

iii. Sampling

One of the methods used for data reduction is sampling as it is capable to reduce the large data set into a much smaller data sample.

DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

iii. Sampling

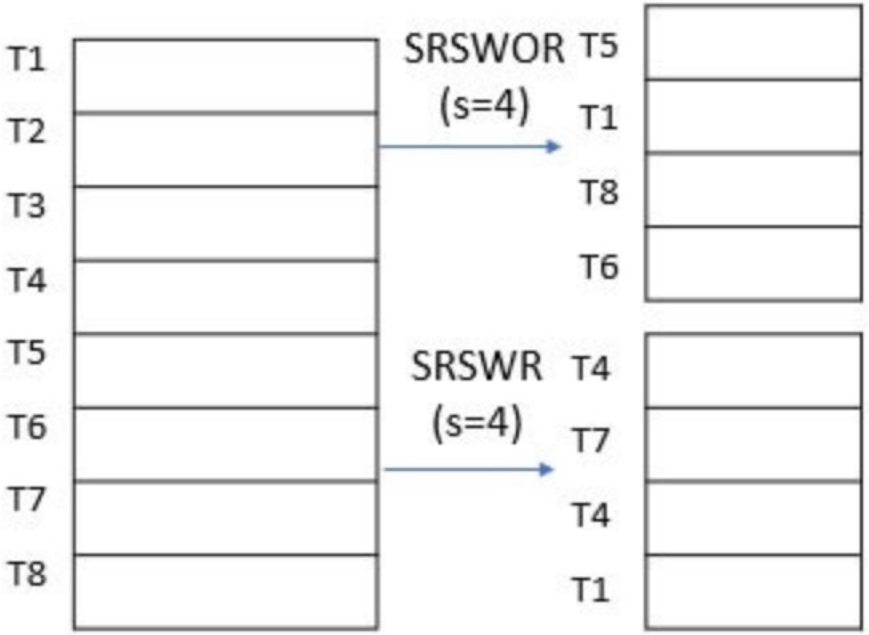
- **Simple random sample without replacement (SRSWOR) of size s:** In this 's number' of tuples are drawn from N tuples such that in the data set D (s<N). The probability of drawing any tuple from the data set D is 1/N this means all tuples have an equal probability of getting sampled.
- **Simple random sample with replacement (SRSWR) of size s:** It is similar to the SRSWOR but the tuple is drawn from data set D, is recorded and then replaced back into the data set D so that it can be drawn again.

DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

iii. Sampling



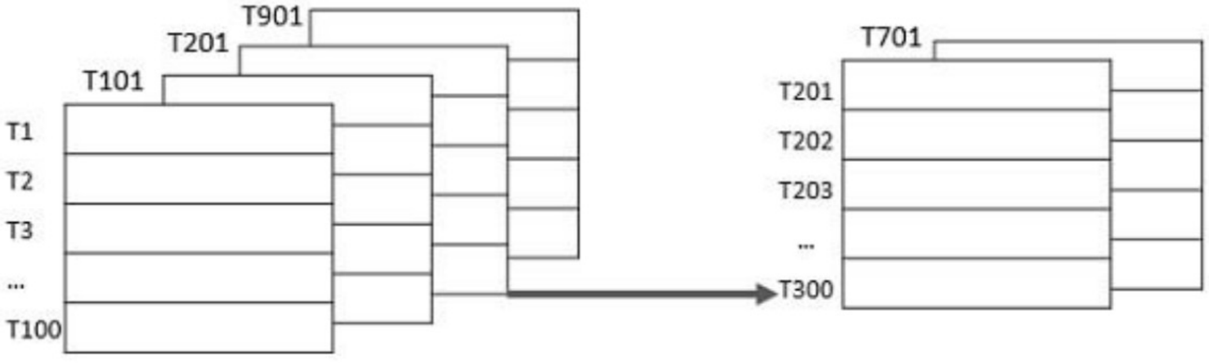
DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

iii. Sampling

Cluster sample: The tuples in data set D are clustered into M mutually disjoint subsets. From these clusters, a simple random sample of size s could be generated where s<M. The data reduction can be applied by implementing SRSWOR on these clusters



Cluster Sample
(s=2)

DATA PREPROCESSING

3. DATA REDUCTION

b. Numerosity Reduction

iii. Sampling

Cluster sample: The tuples in data set D are clustered into M mutually disjoint subsets. From these clusters, a simple random sample of size s could be generated where s<M. The data reduction can be applied by implementing SRSWOR on these clusters