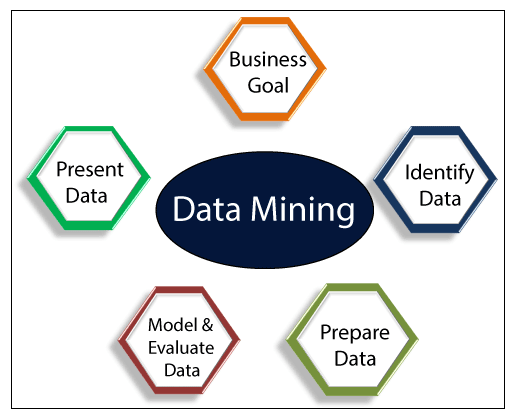
UNIT – I

Introduction to Data Mining

What is Data Mining?

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective. This process includes various types of **services such as text mining, web mining, audio and video mining, pictorial data** mining, and social media mining. It is done through software that is simple or highly specific. By outsourcing data mining, all the work can be done faster with low operation costs. Specialized firms can also use new technologies to collect data that is impossible to locate manually. There are tonnes of information available on various platforms, but very little knowledge is accessible. The biggest challenge is to analyze the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find better insight from it.



Types of Data Mining

Data mining can be performed on the following types of data:

**Relational Database:**

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.

**Data warehouses:**

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

**Data Repositories:**

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

**Object-Relational Database:**

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

**Transactional Database:**

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Advantages of Data Mining

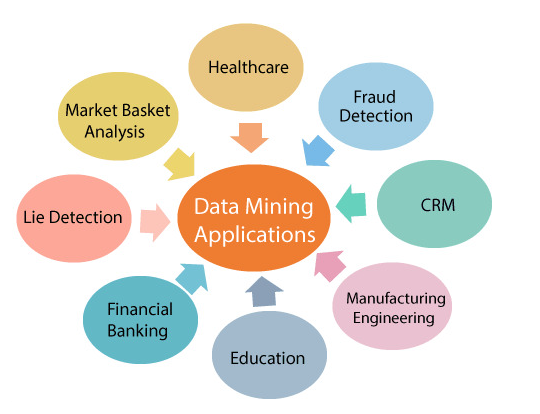
* The Data Mining technique enables organizations to obtain knowledge-based data.
* Data mining enables organizations to make lucrative modifications in operation and production.
* Compared with other statistical data applications, data mining is a cost-efficient.
* Data Mining helps the decision-making process of an organization.
* It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
* It can be induced in the new system as well as the existing platforms.
* It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

* There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
* Many data mining analytics software is difficult to operate and needs advance training to work on.
* Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
* The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

## Data Mining Applications

Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.

`

**Data Mining in Healthcare:**

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

**Data Mining in Market Basket Analysis:**

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

**Data mining in Education:**

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

**Data Mining in Manufacturing Engineering:**

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

**Data Mining in CRM (Customer Relationship Management):**

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

**Data Mining in Fraud detection:**

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

**Data Mining in Lie Detection:**

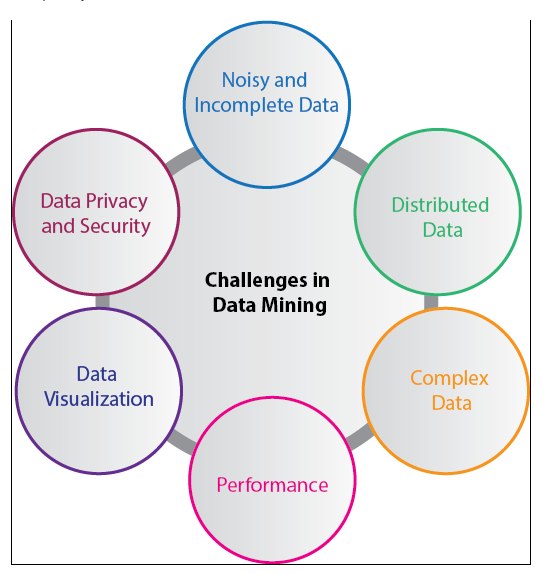
Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

**Data Mining Financial Banking:**

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

## Challenges of Implementation in Data mining

Although data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to performance, data, methods, and techniques, etc. The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.



**Incomplete and noisy data:**

The process of extracting useful data from large volumes of data is data mining. The data in the real-world is heterogeneous, incomplete, and noisy. Data in huge quantities will usually be inaccurate or unreliable. These problems may occur due to data measuring instrument or because of human errors. Suppose a retail chain collects phone numbers of customers who spend more than $ 500, and the accounting employees put the information into their system. The person may make a digit mistake when entering the phone number, which results in incorrect data. Even some customers may not be willing to disclose their phone numbers, which results in incomplete data. The data could get changed due to human or system error. All these consequences (noisy and incomplete data)makes data mining challenging.

**Data Distribution:**

Real-worlds data is usually stored on various platforms in a distributed computing environment. It might be in a database, individual systems, or even on the internet. Practically, It is a quite tough task to make all the data to a centralized data repository mainly due to organizational and technical concerns. For example, various regional offices may have their servers to store their data. It is not feasible to store, all the data from all the offices on a central server. Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.

**Complex Data:**

Real-world data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Managing these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

**Performance:**

The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

**Data Privacy and Security:**

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyzes the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

**Data Visualization:**

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.

*There are many more challenges in data mining in addition to the problems above-mentioned. More problems are disclosed as the actual data mining process begins, and the success of data mining relies on getting rid of all these difficulties.*

Prerequisites

Before learning the concepts of Data Mining, you should have a basic understanding of Statistics, Database Knowledge, and Basic programming language.

Audience

Our Data Mining Tutorial is prepared for all beginners or computer science graduates to help them learn the basics to advanced techniques related to data mining.

Problems

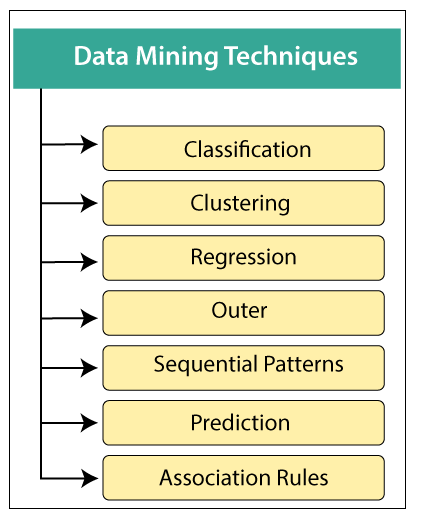
We assure you that you will not find any difficulty while learning our Data Mining tutorial. But if there is any mistake in this tutorial, kindly post the problem or error in the contact form so that we can improve it.

# **Data Mining Techniques**

Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets. These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees. Thus, data mining incorporates analysis and prediction.

Depending on various methods and technologies from the intersection of machine learning, database management, and statistics, professionals in data mining have devoted their careers to better understanding how to process and make conclusions from the huge amount of data, but what are the methods they use to make it happen?

In recent data mining projects, various major data mining techniques have been developed and used, including association, classification, clustering, prediction, sequential patterns, and regression.



1. Classification:

This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

Data mining techniques can be classified by different criteria, as follows:

1. **Classification of Data mining frameworks as per the type of data sources mined:**  
   This classification is as per the type of data handled. For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on..
2. **Classification of data mining frameworks as per the database involved:**  
   This classification based on the data model involved. For example. Object-oriented database, transactional database, relational database, and so on..
3. **Classification of data mining frameworks as per the kind of knowledge discovered:**  
   This classification depends on the types of knowledge discovered or data mining functionalities. For example, discrimination, classification, clustering, characterization, etc. some frameworks tend to be extensive frameworks offering a few data mining functionalities together..
4. **Classification of data mining frameworks according to data mining techniques used:**  
   This classification is as per the data analysis approach utilized, such as neural networks, machine learning, genetic algorithms, visualization, statistics, data warehouse-oriented or database-oriented, etc.  
   The classification can also take into account, the level of user interaction involved in the data mining procedure, such as query-driven systems, autonomous systems, or interactive exploratory systems.

2. Clustering:

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modeling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. From a machine learning point of view, clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent framework represents a data concept. From a practical point of view, clustering plays an extraordinary job in data mining applications. For example, scientific data exploration, text mining, information retrieval, spatial database applications, CRM, Web analysis, computational biology, medical diagnostics, and much more.

In other words, we can say that Clustering analysis is a data mining technique to identify similar data. This technique helps to recognize the differences and similarities between the data. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

3. Regression:

Regression analysis is the data mining process is used to identify and analyze the relationship between variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modeling. For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand, and competition. Primarily it gives the exact relationship between two or more variables in the given data set.

4. Association Rules:

This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set.

Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases. Association rule mining has several applications and is commonly used to help sales correlations in data or medical data sets.

The way the algorithm works is that you have various data, For example, a list of grocery items that you have been buying for the last six months. It calculates a percentage of items being purchased together.

These are three major measurements technique:

* **Lift:**  
  This measurement technique measures the accuracy of the confidence over how often item B is purchased.  
                    **(Confidence) / (item B)/ (Entire dataset)**
* **Support:**  
  This measurement technique measures how often multiple items are purchased and compared it to the overall dataset.  
                    **(Item A + Item B) / (Entire dataset)**
* **Confidence:**  
  This measurement technique measures how often item B is purchased when item A is purchased as well.  
                    **(Item A + Item B)/ (Item A)**

5. Outer detection:

This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behavior. This technique may be used in various domains like intrusion, detection, fraud detection, etc. It is also known as Outlier Analysis or Outilier mining. The outlier is a data point that diverges too much from the rest of the dataset. The majority of the real-world datasets have an outlier. Outlier detection plays a significant role in the data mining field. Outlier detection is valuable in numerous fields like network interruption identification, credit or debit card fraud detection, detecting outlying in wireless sensor network data, etc.

6. Sequential Patterns:

The sequential pattern is a data mining technique specialized for **evaluating sequential data** to discover sequential patterns. It comprises of finding interesting subsequences in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc.

In other words, this technique of data mining helps to discover or recognize similar patterns in transaction data over some time.

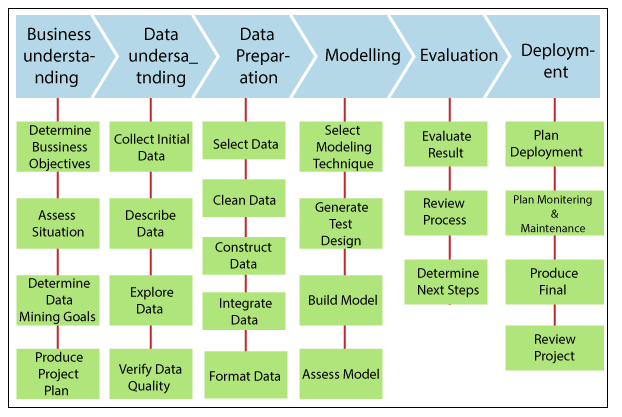
7. Prediction:

Prediction used a combination of other data mining techniques such as trends, clustering, classification, etc. It analyzes past events or instances in the right sequence to predict a future event.

# **Data Mining Implementation Process**

Many different sectors are taking advantage of data mining to boost their business efficiency, including manufacturing, chemical, marketing, aerospace, etc. Therefore, the need for a conventional data mining process improved effectively. Data mining techniques must be reliable, repeatable by company individuals with little or no knowledge of the data mining context. As a result, a cross-industry standard process for data mining (CRISP-DM) was first introduced in 1990, after going through many workshops, and contribution for more than 300 organizations.

Data mining is described as a pro0cess of finding hidden precious data by evaluating the huge quantity of information stored in data warehouses, using multiple data mining techniques such as Artificial Intelligence (AI), Machine learning and statistics.



# **Data Objects and Attributes and Quality:**

**Data:** It is how the data objects and their attributes are stored.

* An **attribute** is an object’s property or characteristics. For example. A person’s hair colour, air humidity etc.
* An attribute set defines an **object**. The **object** is also referred to as a record of the instances or entity.

Different **types of attributes or data types:**

1. **Nominal Attribute:**   
   Nominal Attributes only provide enough attributes to differentiate between one object and another. Such as Student Roll No., Sex of the Person.
2. **Ordinal Attribute:**   
   The ordinal attribute value provides sufficient information to order the objects. Such as Rankings, Grades, Height
3. **Binary Attribute:**   
   These are 0 and 1. Where 0 is the absence of any features and 1 is the inclusion of any characteristics.
4. **Numeric attribute:**It is quantitative, such that quantity can be measured and represented in integer or real values ,are of two types  
   **Interval Scaled attribute:**   
   It is measured on a scale of equal size units,these attributes allow us to compare such as temperature in C or F and thus values of attributes have ordered.  
    **Ratio Scaled attribute:**   
   Both differences and ratios are significant for Ratio. For eg. age, length, and Weight.

**Data Quality: Why do we preprocess the data?**   
Many characteristics act as a deciding factor for data quality, such as incompleteness and incoherent information, which are common properties of the big database in the real world. Factors used for data quality assessment are:

* **Accuracy:**   
  There are many possible reasons for flawed or inaccurate data here. i.e. Having incorrect values of properties that could be human or computer errors.
* **Completeness:**   
  For some reasons, incomplete data can occur, attributes of interest such as customer information for sales & transaction data may not always be available.
* **Consistency:**   
  Incorrect data can also result from inconsistencies in naming convention or data codes, or from input field incoherent format. Duplicate tuples need cleaning of details, too.
* **Timeliness:**   
  It also affects the quality of the data. At the end of the month, several sales representatives fail to file their sales records on time. There are also several corrections & adjustments which flow into after the end of the month. Data stored in the database are incomplete for a time after each month.
* **Believability:**   
  It is reflective of how much users trust the data.
* **Interpretability:**   
  It is a reflection of how easy the users can understand the data.

# **Basic Statistical Methods in Data Mining:**

Data mining refers to extracting or mining knowledge from large amounts of data. In other words, data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns. Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective, and accurate. Any situation can be analyzed in two ways in data mining:

* **Statistical Analysis:** In statistics, data is collected, analyzed, explored, and presented to identify patterns and trends. Alternatively, it is referred to as quantitative analysis.
* **Non-statistical Analysis:** This analysis provides generalized information and includes sound, still images, and moving images.

In statistics, there are two main categories:

* **Descriptive Statistics:** The purpose of descriptive statistics is to organize data and identify the main characteristics of that data. Graphs or numbers summarize the data. Average, Mode, SD(Standard Deviation), and Correlation are some of the commonly used descriptive statistical methods.
* **Inferential Statistics:** The process of drawing conclusions based on probability theory and generalizing the data. By analyzing sample statistics, you can infer parameters about populations and make models of relationships within data.

There are various statistical terms that one should be aware of while dealing with statistics. Some of these are:

* Population
* Sample
* Variable
* Quantitative Variable
* Qualitative Variable
* Discrete Variable
* Continuous Variable

Now, let’s start discussing statistical methods. This is the analysis of raw data using mathematical formulas, models, and techniques. Through the use of statistical methods, information is extracted from research data, and different ways are available to judge the robustness of research outputs.

As a matter of fact, today’s statistical methods used in the data mining field typically are derived from the vast statistical toolkit developed to answer problems arising in other fields. These techniques are taught in science curriculums. It is necessary to check and test several hypotheses. The hypotheses described above help us assess the validity of our data mining endeavor when attempting to infer any inferences from the data under study. When using more complex and sophisticated statistical estimators and tests, these issues become more pronounced.

For extracting knowledge from databases containing different types of observations, a variety of statistical methods are available in Data Mining and some of these are:

* Logistic regression analysis
* Correlation analysis
* Regression analysis
* Discriminate analysis
* Linear discriminant analysis (LDA)
* Classification
* Clustering
* Outlier detection
* Classification and regression trees,
* Correspondence analysis
* Nonparametric regression,
* Statistical pattern recognition,
* Categorical data analysis,
* Time-series methods for trends and periodicity
* Artificial neural networks

Now, let’s try to understand some of the important statistical methods which are used in data mining:

* **Linear Regression:** The linear regression method uses the best linear relationship between the independent and dependent variables to predict the target variable. In order to achieve the best fit, make sure that all the distances between the shape and the actual observations at each point are as small as possible. A good fit can be determined by determining that no other position would produce fewer errors given the shape chosen. Simple linear regression and multiple linear regression are the two major types of linear regression. By fitting a linear relationship to the independent variable, the simple linear regression predicts the dependent variable. Using multiple independent variables, multiple linear regression fits the best linear relationship with the dependent variable. For more details, you can refer [linear regression.](https://www.geeksforgeeks.org/linear-regression-python-implementation/#:~:text=Simple%20linear%20regression%20is%20an,or%20independent%20variable(x).)
* **Classification:** This is a method of data mining in which a collection of data is categorized so that a greater degree of accuracy can be predicted and analyzed. An effective way to analyze very large datasets is to classify them. Classification is one of several methods aimed at improving the efficiency of the analysis process. A Logistic Regression and a Discriminant Analysis stand out as two major classification techniques.
  + **Logistic Regression:** It can also be applied to machine learning applications and predictive analytics. In this approach, the dependent variable is either binary (binary regression) or multinomial (multinomial regression): either one of the two or a set of one, two, three, or four options. With a logistic regression equation, one can estimate probabilities regarding the relationship between the independent variable and the dependent variable. For understanding logistic regression analysis in detail, you can refer to logistic regression.
  + **Discriminant Analysis:** A Discriminant Analysis is a statistical method of analyzing data based on the measurements of categories or clusters and categorizing new observations into one or more populations that were identified a priori. The discriminant analysis models each response class independently then uses Bayes’s theorem to flip these projections around to estimate the likelihood of each response category given the value of X. These models can be either linear or quadratic.
    - **Linear Discriminant Analysis:**  According to **Linear Discriminant Analysis,** each observation is assigned a discriminant score to classify it into a response variable class. By combining the independent variables in a linear fashion, these scores can be obtained. Based on this model, observations are drawn from a Gaussian distribution, and the predictor variables are correlated across all k levels of the response variable, Y. and for further details [linear discriminant analysis](https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/)
    - **Quadratic Discriminant Analysis:**  An alternative approach is provided by Quadratic Discriminant Analysis. LDA and QDA both assume Gaussian distributions for the observations of the Y classes. Unlike LDA, QDA considers each class to have its own covariance matrix. As a result, the predictor variables have different variances across the k levels in Y.
  + **Correlation Analysis:** In statistical terms, correlation analysis captures the relationship between variables in a pair. The value of such variables is usually stored in a column or rows of a database table and represents a property of an object.
  + **Regression Analysis:** Based on a set of numeric data, regression is a data mining method that predicts a range of numerical values (also known as continuous values). You could, for instance, use regression to predict the cost of goods and services based on other variables. A regression model is used across numerous industries for forecasting financial data, modeling environmental conditions, and analyzing trends.

The first step in creating good statistics is having good data that was derived with an aim in mind. There are two main types of data: an input (independent or predictor) variable, which we control or are able to measure, and an output (dependent or response) variable which is observed. A few will be quantitative measurements, but others may be qualitative or categorical variables (called factors).

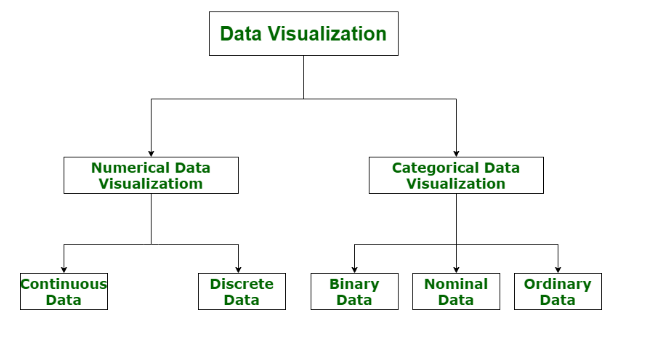
# **Data Visualization**

**Data visualization** is actually a set of data points and information that are represented graphically to make it easy and quick for user to understand. Data visualization is good if it has a clear meaning, purpose, and is very easy to interpret, without requiring context. Tools of data visualization provide an accessible way to see and understand trends, outliers, and patterns in data by using visual effects or elements such as a chart, graphs, and maps.

**Characteristics of Effective Graphical Visual :**

* It shows or visualizes data very clearly in an understandable manner.
* It encourages viewers to compare different pieces of data.
* It closely integrates statistical and verbal descriptions of data set.
* It grabs our interest, focuses our mind, and keeps our eyes on message as human brain tends to focus on visual data more than written data.
* It also helps in identifying area that needs more attention and improvement.
* Using graphical representation, a story can be told more efficiently. Also, it requires less time to understand picture than it takes to understand textual data.

**Categories of Data Visualization ;**  
Data visualization is very critical to market research where both numerical and categorical data can be visualized that helps in an increase in impacts of insights and also helps in reducing risk of analysis paralysis. So, data visualization is categorized into following categories :



**Figure –** Categories of Data Visualization

1. **Numerical Data :**  
   Numerical data is also known as Quantitative data. Numerical data is any data where data generally represents amount such as height, weight, age of a person, etc. Numerical data visualization is easiest way to visualize data. It is generally used for helping others to digest large data sets and raw numbers in a way that makes it easier to interpret into action. Numerical data is categorized into two categories :
   * **Continuous Data –**  
     It can be narrowed or categorized (Example: Height measurements).
   * **Discrete Data –**  
     This type of data is not “continuous” (Example: Number of cars or children’s a household has).

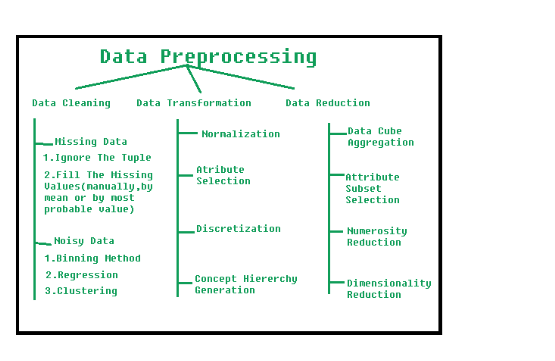
The type of visualization techniques that are used to represent numerical data visualization is Charts and Numerical Values. Examples are Pie Charts, Bar Charts, Averages, Scorecards, etc.

1. **Categorical Data :**  
   Categorical data is also known as Qualitative data. Categorical data is any data where data generally represents groups. It simply consists of categorical variables that are used to represent characteristics such as a person’s ranking, a person’s gender, etc. Categorical data visualization is all about depicting key themes, establishing connections, and lending context. Categorical data is classified into three categories :
   * **Binary Data –**  
     In this, classification is based on positioning (Example: Agrees or Disagrees).
   * **Nominal Data –**  
     In this, classification is based on attributes (Example: Male or Female).
   * **Ordinal Data –**  
     In this, classification is based on ordering of information (Example: Timeline or processes).

The type of visualization techniques that are used to represent categorical data is Graphics, Diagrams, and Flowcharts. Examples are Word clouds, Sentiment Mapping, Venn Diagram, etc.

# **Data Pre-processing in Data Mining**

**Preprocessing in Data Mining:**   
Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



**Steps Involved in 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.   
  Some of them are:
  1. **Ignore the tuples:**   
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **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.
* **(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 :
  1. **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. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**   
     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**   
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**   
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**   
   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)
2. **Attribute Selection:**   
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**   
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**   
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**   
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**   
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**   
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**   
   This enable to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**   
   This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are:Wavelet transforms and PCA (Principal Component Analysis).

# **Data Cleaning in Data Mining**

Data cleaning is a crucial process in Data Mining. It carries an important part in the building of a model. Data Cleaning can be regarded as the process needed, but everyone often neglects it. Data quality is the main issue in quality information management. Data quality problems occur anywhere in information systems. These problems are solved by data cleaning.

Data cleaning is fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

Generally, data cleaning reduces errors and improves data quality. Correcting errors in data and eliminating bad records can be a time-consuming and tedious process, but it cannot be ignored. Data mining is a key technique for data cleaning. Data mining is a technique for discovering interesting information in data. Data quality mining is a recent approach applying data mining techniques to identify and recover data quality problems in large databases. Data mining automatically extracts hidden and intrinsic information from the collections of data. Data mining has various techniques that are suitable for data cleaning.

### **Steps of Data Cleaning**

While the techniques used for data cleaning may vary according to the types of data your company stores, you can follow these basic steps to cleaning your data, such as:

**1. Remove duplicate or irrelevant observations**

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze.

For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient, minimize distraction from your primary target, and create a more manageable and performable dataset.

**2. Fix structural errors**

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes. For example, you may find "N/A" and "Not Applicable" in any sheet, but they should be analyzed in the same category.

**3. Filter unwanted outliers**

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data entry, doing so will help the performance of the data you are working with.

However, sometimes, the appearance of an outlier will prove a theory you are working on. And just because an outlier exists doesn't mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

**4. Handle missing data**

You can't ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered, such as:

* You can drop observations with missing values, but this will drop or lose information, so be careful before removing it.
* You can input missing values based on other observations; again, there is an opportunity to lose the integrity of the data because you may be operating from assumptions and not actual observations.
* You might alter how the data is used to navigate null values effectively.

**5. Validate data**

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation, such as:

* Does the data make sense?
* Does the data follow the appropriate rules for its field?
* Does it prove or disprove your working theory or bring any insight to light?
* Can you find trends in the data to help you for your next theory?
* If not, is that because of a data quality issue?

Because of incorrect or noisy data, false conclusions can inform poor business strategy and decision-making. False conclusions can lead to an embarrassing moment in a reporting meeting when you realize your data doesn't stand up to study. Before you get there, it is important to create a culture of quality data in your organization. To do this, you should document the tools you might use to create this strategy.

### **Methods of Data Cleaning**

There are many data cleaning methods through which the data should be run. The methods are described below:

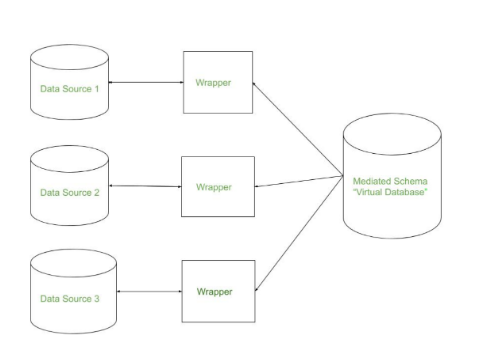


1. **Ignore the tuples:** This method is not very feasible, as it only comes to use when the tuple has several attributes is has missing values.
2. **Fill the missing value:** This approach is also not very effective or feasible. Moreover, it can be a time-consuming method. In the approach, one has to fill in the missing value. This is usually done manually, but it can also be done by attribute mean or using the most probable value.
3. **Binning method:** This approach is very simple to understand. The smoothing of sorted data is done using the values around it. The data is then divided into several segments of equal size. After that, the different methods are executed to complete the task.
4. **Regression:** The data is made smooth with the help of using the regression function. The regression can be linear or multiple. Linear regression has only one independent variable, and multiple regressions have more than one independent variable.
5. **Clustering:** This method mainly operates on the group. Clustering groups the data in a cluster. Then, the outliers are detected with the help of clustering. Next, the similar values are then arranged into a "group" or a "cluster".

# **Data Integration in Data Mining**

**Data Integration** is a data preprocessing technique that combines data from multiple heterogeneous data sources into a coherent data store and provides a unified view of the data. These sources may include multiple data cubes, databases, or flat files.

The data integration approaches are formally defined as triple <G, S, M> where,   
G stand for the global schema,   
S stands for the heterogeneous source of schema,   
M stands for mapping between the queries of source and global schema.



There are mainly 2 major approaches for data integration – one is the “tight coupling approach” and another is the “loose coupling approach”.

**Tight Coupling:**

* Here, a data warehouse is treated as an information retrieval component.
* In this coupling, data is combined from different sources into a single physical location through the process of ETL – Extraction, Transformation, and Loading.

**Loose Coupling:**

* Here, an interface is provided that takes the query from the user, transforms it in a way the source database can understand, and then sends the query directly to the source databases to obtain the result.
* And the data only remains in the actual source databases.

# **Data Reduction in Data Mining**

The method of data reduction may achieve a condensed description of the original data which is much smaller in quantity but keeps the quality of the original data.

**Methods of data reduction:**   
These are explained as following below.

**1. Data Cube Aggregation:**   
This technique is used to aggregate data in a simpler form. For example, imagine the information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months. They involve you in the annual sales, rather than the quarterly average,  So we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.

**2. Dimension reduction:**   
Whenever we come across any data which is weakly important, then we use the attribute required for our analysis. It reduces data size as it eliminates outdated or redundant features.

* **Step-wise Forward Selection –**   
  The selection begins with an empty set of attributes later on we decide the best of the original attributes on the set based on their relevance to other attributes. We know it as a p-value in statistics.

Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}

Initial reduced attribute set: { }

Step-1: {X1}

Step-2: {X1, X2}

Step-3: {X1, X2, X5}

Final reduced attribute set: {X1, X2, X5}

* **Step-wise Backward Selection –**   
  This selection starts with a set of complete attributes in the original data and at each point, it eliminates the worst remaining attribute in the set.

Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}

Initial reduced attribute set: {X1, X2, X3, X4, X5, X6 }

Step-1: {X1, X2, X3, X4, X5}

Step-2: {X1, X2, X3, X5}

Step-3: {X1, X2, X5}

Final reduced attribute set: {X1, X2, X5}

* **Combination of forwarding and Backward Selection –**   
  It allows us to remove the worst and select the best attributes, saving time and making the process faster.

**3.**[**Data Compression**](https://www.geeksforgeeks.org/difference-between-lossy-compression-and-lossless-compression/)**:**   
The data compression technique reduces the size of the files using different encoding mechanisms (Huffman Encoding & run-length Encoding). We can divide it into two types based on their compression techniques.

* **Lossless Compression –**   
  Encoding techniques (Run Length Encoding) allow a simple and minimal data size reduction. Lossless data compression uses algorithms to restore the precise original data from the compressed data.
* **Lossy Compression –**   
  Methods such as the Discrete Wavelet transform technique, PCA (principal component analysis) are examples of this compression. For e.g., the JPEG image format is a lossy compression, but we can find the meaning equivalent to the original image. In lossy-data compression, the decompressed data may differ from the original data but are useful enough to retrieve information from them.

**4. [Numerosity Reduction](https://www.geeksforgeeks.org/numerosity-reduction-in-data-mining/):**   
In this reduction technique, the actual data is replaced with mathematical models or smaller representations of the data instead of actual data, it is important to only store the model parameter. Or non-parametric methods such as clustering, histogram, and sampling.

**5. Discretization & Concept Hierarchy Operation:**   
Techniques of data discretization are used to divide the attributes of the continuous nature into data with intervals. We replace many constant values of the attributes by labels of small intervals. This means that mining results are shown in a concise, and easily understandable way.

* **Top-down discretization –**   
  If you first consider one or a couple of points (so-called breakpoints or split points) to divide the whole set of attributes and repeat this method up to the end, then the process is known as top-down discretization also known as splitting.
* **Bottom-up discretization –**   
  If you first consider all the constant values as split points, some are discarded through a combination of the neighborhood values in the interval, that process is called bottom-up discretization.

# **Data Transformation in Data Mining**

The data are transformed in ways that are ideal for mining the data. The data transformation involves steps that are:

**1. Smoothing:**  
It is a process that is used to remove noise from the dataset using some algorithms It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.

The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn’t see otherwise.

**2. Aggregation:**  
Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.

The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies.

For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

**3. Discretization:**  
It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes.

Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.

For **example**, (1-10, 11-20) (age:- young, middle age, senior).

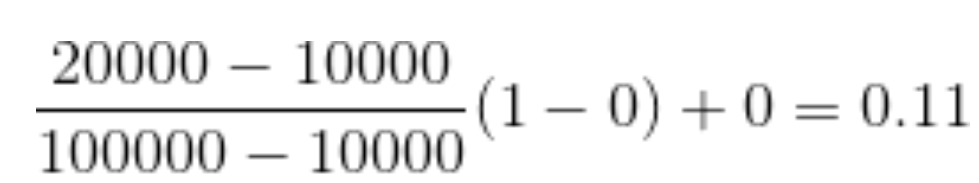
**4. Attribute Construction:**  
Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.

**5. Generalization:**  
It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).

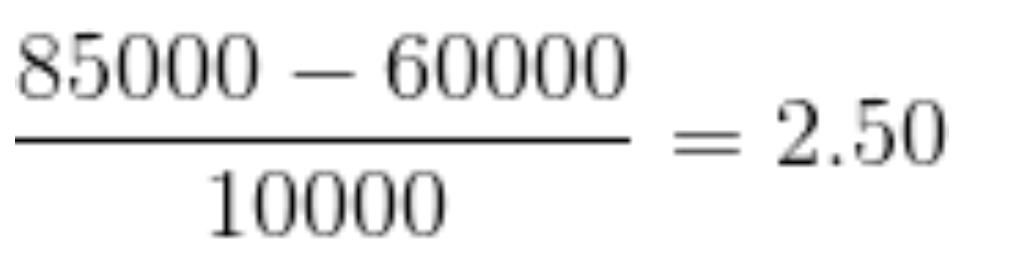
For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

**6. Normalization:** Data normalization involves converting all data variable into a given range.  
Techniques that are used for normalization are:

* **Min-Max Normalization:**
  + This transforms the original data linearly.
  + Suppose that: min\_A is the minima and max\_A is the maxima of an attribute, P

We Have the Formula:  


* + Where v is the value you want to plot in the new range.
  + v’ is the new value you get after normalizing the old value.

Solved **example**:  
Suppose the minimum and maximum value for an attribute profit(P) are Rs. 10, 000 and Rs. 100, 000. We want to plot the profit in the range [0, 1]. Using min-max normalization the value of Rs. 20, 000 for attribute profit can be plotted to:  


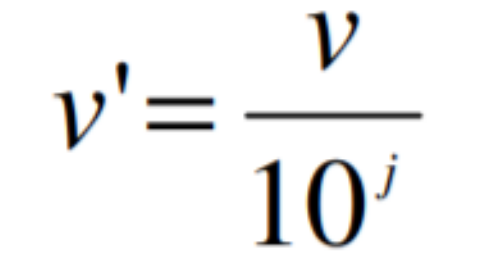
And hence, we get the value of v’ as 0.11

* **Z-Score Normalization:**
  + In z-score normalization (or zero-mean normalization) the values of an attribute (A), are normalized based on the mean of A and its standard deviation
  + A value, v, of attribute A is normalized to v’ by computing

For **example**:  
Let mean of an attribute P = 60, 000, Standard Deviation = 10, 000, for the attribute P. Using z-score normalization, a value of 85000 for P can be transformed to:  


And hence we get the value of v’ to be 2.5

* **Decimal Scaling:**
  + It normalizes the values of an attribute by changing the position of their decimal points
  + The number of points by which the decimal point is moved can be determined by the absolute maximum value of attribute A.
  + A value, v, of attribute A is normalized to v’ by computing



* + where j is the smallest integer such that Max(|v’|) < 1.

For **example**:

* + Suppose: Values of an attribute P varies from -99 to 99.
  + The maximum absolute value of P is 99.
  + For normalizing the values we divide the numbers by 100 (i.e., j = 2) or (number of integers in the largest number) so that values come out to be as 0.98, 0.97 and so on.

# **Data discretization in data mining (or)**

**Data De centralization in Data Mining:**

Data discretization converts a large number of data values into smaller once, so that data evaluation and data management becomes very easy.

## Data discretization example

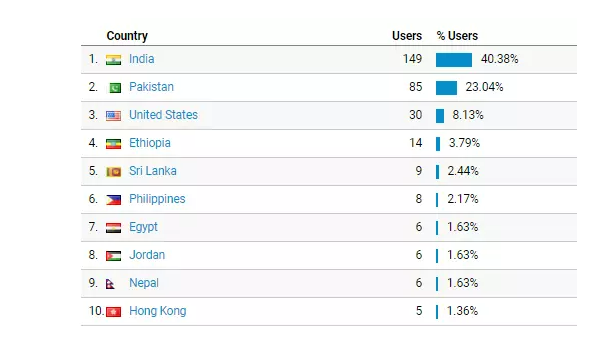
we have an attribute of age with the following values.

|  |  |
| --- | --- |
| **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 |

Another example is the Website visitor’s data. As seen in the figure below, data is discretized into the countries.



All IP addresses of a specific country are discretized into the countries.

For example, all visitors visit the website with the IP addresses of the United States are shown under country labels.

## What are some famous techniques of data discretization?

1. Histogram analysis: **Histogram** is a plot used to present the underlying frequency distribution of a set of continuous data. The histogram helps the inspection of the data for the distribution of the data. For example normal distribution representation, outliers, and skewness representation, etc.
2. Binning: **Binning** is a data smoothing technique and its helps to group a huge number of continuous values into a smaller number of bins. For example, if we have data about a group of students, and we want to arrange their marks into a smaller number of marks intervals by making the bins of grades. One bin for grade A, one for grade B, one for C, one for D, and one for F Grade.
3. **Correlation analysis:**Cluster analysis is commonly known as [clustering](https://t4tutorials.com/correlation-analysis-of-numerical-data-in-data-mining/). Clustering is the task of grouping similar objects in one group, commonly called clusters.  All different objects are placed in different clusters.
4. Clustering analysis
5. [Decision tree](https://t4tutorials.com/decision-tree-induction-calculation-on-categorical-attributes-in-data-mining/) analysis
6. Equal width partitioning
7. Equal depth partitioning