ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

JAGAT GURU NANAK DEV
PUNJAB STATE OPEN UNIVERSITY, PATIALA

(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

The Motto of the University
(SEWA)

**SKILL ENHANCEMENT EMPLOYABILITY WISDOM
ACCESSIBILITY**

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ, ਪਟਿਆਲਾ
JAGAT GURU NANAK DEV
PUNJAB STATE OPEN UNIVERSITY, PATIALA

M.SC. (COMPUTER SCIENCE)
SEMESTER-III
COURSE: DATA MINING & VISUALIZATION

COURSE CODE: MSCS-3-02T

**ADDRESS: C/28, THE LOWER MALL, PATIALA-147001
WEBSITE: www.psou.ac.in**

M.Sc. (Computer Science)
Semester-3
MSCS-3-02T: Data Mining and Visualization

Total Marks: 100
External Marks: 70
Internal Marks: 30
Credits: 4
Pass Percentage: 40%

INSTRUCTIONS FOR THE PAPER SETTER/EXAMINER

1. The syllabus prescribed should be strictly adhered to.
2. The question paper will consist of three sections: A, B, and C. Sections A and B will have four questions from the respective sections of the syllabus and will carry 10 marks each. The candidates will attempt two questions from each section.
3. Section C will have fifteen short answer questions covering the entire syllabus. Each question will carry 3 marks. Candidates will attempt any ten questions from this section.
4. The examiner shall give a clear instruction to the candidates to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.
5. The duration of each paper will be three hours.

INSTRUCTIONS FOR THE CANDIDATES

Candidates are required to attempt any two questions each from the sections A and B of the question paper and any ten short q questions from Section C. They have to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.

Course: Data Mining & Visualization	
Course Code: MSCS-3-02T	
Course Outcomes (COs) After the completion of this course, the students will be able to:	
CO1	Understand Data Warehouse fundamentals and Data Mining tools.
CO2	Understand Data Mining Techniques
CO3	Apply clustering methods like K means, hierarchical clustering, agglomerative clustering, divisive clustering to solve problems and evaluate clusters
CO4	Gain knowledge related to application areas of data mining
CO5	Understand the components involved in data visualization design.

SECTION-A

Unit 1: Data Mining: Introduction, Scope of Data Mining; How does Data Mining Works, Predictive Modeling: Data Mining and Data Warehousing: Architecture for Data Mining: Profitable Applications: Data Mining Tools.

Unit II: Data Pre-processing: Overview, Data Cleaning, Data Integration and Transformation, Data Reduction, Discretization and Concept Hierarchy Generation.

Unit III: Data Mining Techniques: An Overview, Data Mining Versus Database Management System, Data Mining Techniques- Association rules, Classification, Regression, Clustering, Neural networks.

Unit IV: Clustering: Introduction, Cluster Analysis, Clustering Methods- K means, Hierarchical clustering, Agglomerative clustering, Divisive clustering, evaluating clusters.

SECTION-B

Unit V: Applications of Data Mining: Introduction, Business Applications Using Data Mining-Risk management and targeted marketing, Customer profiles and feature construction, Medical applications (diabetic screening), Scientific Applications using Data Mining, Other Applications.

Unit VI: Data Visualization: Introduction, Acquiring and Visualizing Data, Simultaneous acquisition and visualization, Applications of Data Visualization, Keys factors of Data Visualization (Control of Presentation, Faster and Better Java Script processing, Rise of HTML 5, Lowering the implementation Bar).

Unit VII: Exploring the Visual Data Spectrum: Charting Primitives (Data Points, Line Charts, Bar Charts, Pie Charts, Area Charts), Exploring advanced Visualizations (Candlestick Charts, Bubble Charts, Surface Charts, Map Charts, Infographics).

Unit VIII: Visualizing data Programmatically: starting with Google charts (Google Charts API Basics, A Basic bar chart, A basic Pie chart, Working with Chart Animations)

Reference Books:

- Jiawei Han, Micheline Kamber and Jian Pei, “Data Mining Concepts and Techniques”, 3rd Edition, 2000.
- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, “Introduction to Data Mining”, Pearson 2005.
- M. Kantardzic, “Data Mining: Concepts, Models, Methods, and Algorithms”, 2nd edition, Wiley-IEEE Press, 2011.
- Jon Raasch, Graham Murray, Vadim Ogievetsky, Joseph Lowery, “Java Script and j Query for Data Analysis and Visualization”, 2014.

- Ben Fry, “Visualizing data: Exploring and explaining data with the processing environment”, O'Reilly, 2007.

DATA MINING AND VISUALIZATION

UNIT-I DATA MINING

STRUCTURE

1.1 INTRODUCTION

Data mining is knowledge discovery from data. Data mining is also called knowledge discovery and data mining (KDD). Lots of data is being collected and stored. Examples are web data, e-commerce, purchases at names, which persons are the least likely to default on their credit cards? Which types of transactions are likely to be fraudulent, given the transactional history of a particular customer? Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor? Data Mining helps extract such information

Data Explosion Problem

Automated data collection tools and mature database technology leads to tremendous amount of data stored in databases and data warehouses. The fact is that we are drowning in data, but starving for knowledge!

Data mining involves extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. It is the application of descriptive and predictive analysis to support the marketing, sales and service functions. Although data mining can be performed on operational databases, it is more commonly applied to the more stable datasets held in data marts or warehouses

1.2 SCOPE

Data mining technology can generate new business opportunities by providing databases of sufficient size and quality, which leads to extraction of interesting knowledge of rules, regularities, patterns, and constraints in data in large data bases.

Automated prediction of trends and behaviors: Data mining automates the process of finding predictive information in large databases. Predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automated discovery of previously unknown patterns: Data mining tools inspect databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together.

1.3 WHAT IS DATA MINING

Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining is the extraction of hidden predictive information from large databases and is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They examine databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

What kind of data can be mined?

- Database-oriented data sets and applications
- Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
- Data streams and sensor data
- Time-series data, temporal data, sequence data (incl. bio-sequences)
- Structure data, graphs, social networks and multi-linked data
- Object-relational databases
- Heterogeneous databases and legacy databases
- Spatial data and spatiotemporal data
- Multimedia database
- Text databases
- The World-Wide Web

1.4 HOW DOES DATA MINING WORKS

Data mining is the process of understanding data through cleaning raw data, finding patterns, creating models, and testing those models. It includes statistics, machine learning, and database systems. After this, application software sorts the data based on the user's results, and finally, the end-user presents the data in an easy-to-understand format, such as a graph or table.

Data mining involves six phase workflow:

The first step in data mining is data collection. Today’s organizations can collect records, logs, website visitors’ data, application data, sales data, and more every day.

1. **Business understanding**

Comprehensive data mining projects start by first identifying project objectives and scope. The business investors will ask a question or state a problem that data mining can answer or solve.

2. Data understanding

Once the business problem is understood, data relevant to the problem is collected. This data often comes from multiple sources, including structured data and unstructured data. This stage may include some investigative analysis to expose some preliminary patterns. At the end of this phase, the data mining team has selected the subset of data for analysis and modeling.

3. Data preparation

This phase begins with more thorough work. Data preparation involves preparing the final data set, which includes all the relevant data needed to answer the business question. Stakeholders will identify the dimensions and variables to explore and prepare the final data set for model creation.

4. Modeling

In this phase, the appropriate modeling techniques are selected for the given data. These techniques can include clustering, predictive models, classification, estimation, or a combination.

5. Evaluation

After creating the models, the test and measure of their success is evaluated. The model may answer things not accounted for, and the model needs to be edited. This phase is designed to allow to look at the progress so far and ensure it's on the right track for meeting the business goals.

6. Deployment

Finally, once the model is accurate and reliable, it is installed it in the real world. The deployment can take place within the organization, be shared with customers, or be used to generate a report for participants to prove its reliability. The work doesn't end when the last line of code is complete; deployment requires careful thought, and a way to make sure the right people are appropriately informed. The data mining team is responsible for the user's understanding of the project.

Data mining involves six common classes of tasks:

1. Anomaly detection – The identification of unusual data records, that might be interesting or data errors that require further investigation.
2. Association rule learning (Dependency modelling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes.
3. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
4. Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
5. Regression – attempts to find a function which models the data with the least error.

6. Summarization – providing a more compact representation of the data set, including visualization and report generation

1.5 PREDICTIVE MODELING

Predictive modeling are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes. Predictive models analyze past performance to assess how likely a customer is to exhibit a specific behavior in the future.

In predictive modeling, data is collected, a statistical model is formulated, predictions are made, and the model is validated as additional data becomes available. For example, risk models can be created to combine member information in complex ways with lifestyle information from external sources to improve assuring accuracy. This category also includes models that seek out indirect data patterns to answer questions about customer performance, such as fraud detection models. Predictive models often perform calculations during live transactions. For example, to evaluate the risk of a given customer or transaction to guide a decision

1.6 DATA MINING AND DATA WAREHOUSING

Data warehousing is the process of pooling all relevant data together. Data mining is considered as a process of extracting data from large data sets.

Data mining comprises the following:

Data mining is the process of analyzing data patterns.

Data is analyzed regularly.

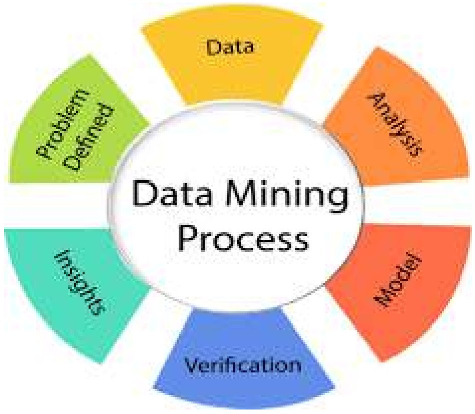
Data mining is the use of pattern recognition logic to identify patterns

Data mining is carried by business users with the help of engineers.

Data mining is considered as a process of extracting data from large data sets.

Data is stored periodically.

Data mining is carried by business users with the help of engineers.



Data Mining Process

Data warehousing comprises the following:

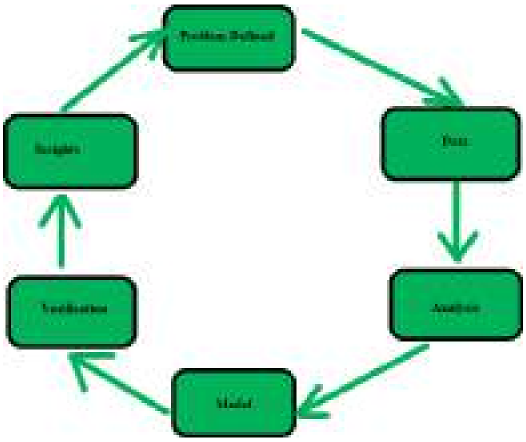
A data warehouse is database system which is designed for analytical analysis instead of transactional work.

Data is stored periodically.

Data warehousing is the process of extracting and storing data to allow easier reporting.

Data warehousing is solely carried out by engineers.

Data warehousing is the process of pooling all relevant data together.



Data warehouse

1.7 ARCHITECTURE FOR DATA MINING

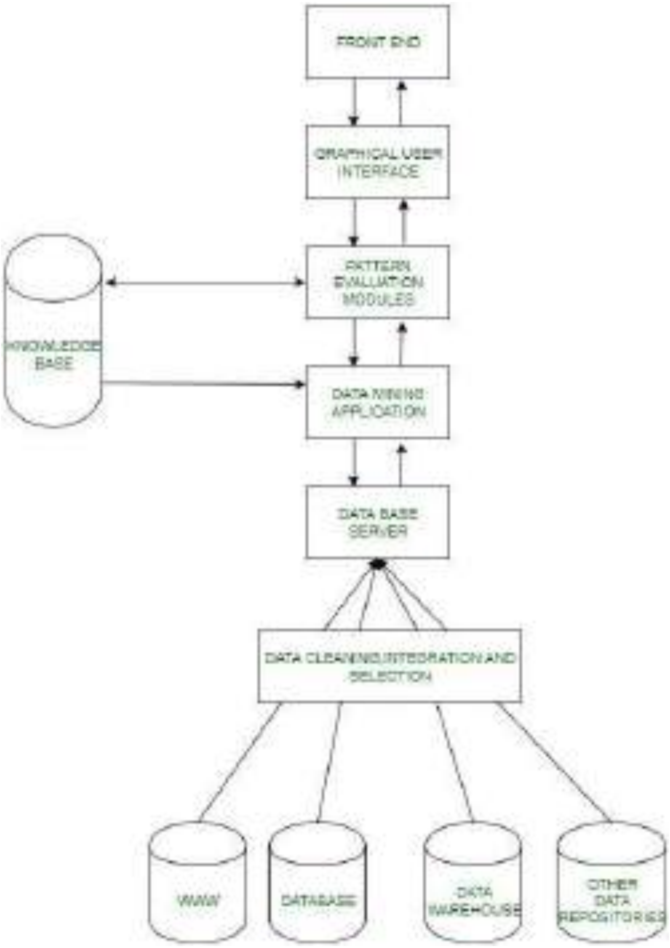
Data mining is an important method where previously unknown and potentially useful information is extracted from the vast amount of data. The data mining process involves several components, and these components constitute a data mining system architecture.

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.

Basic Working:

- It starts when the user puts up certain data mining requests, these requests are then sent to data mining engines for pattern evaluation.
- These applications try to find the solution of the query using the already present database.

- The metadata then extracted is sent for proper analysis to the data mining engine which sometimes interacts with pattern evaluation modules to determine the result.
- This result is then sent to the front end in an easily understandable manner using a suitable interface.



Architecture for Data Mining

The architecture of data mining has the following components:

1. **Data Sources**
   Database, World Wide Web (WWW) and data warehouse are parts of data sources. The data in these sources may be in the form of plain text, spreadsheets or in other forms of media like photos or videos. WWW is one of the biggest sources of data.
2. **Database Server**
   The database server contains the actual data ready to be processed. It performs the task of handling data retrieval as per the request of the user.
3. **Data Mining Engine**
   It is one of the core components of the data mining architecture that performs all kinds of data mining techniques like association, classification, characterization, clustering, prediction, etc.

4. **Pattern Evaluation Modules**
   They are responsible for finding interesting patterns in the data and sometimes they also interact with the database servers for producing the result of the user requests.
5. **Graphic User Interface**
   Since the user cannot fully understand the complexity of the data mining process so graphical user interface helps the user to communicate effectively with the data mining system.
6. **Knowledge Base**
   Knowledge Base is an important part of the data mining engine that is quite beneficial in guiding the search for the result patterns. Data mining engine may also sometimes get inputs from the knowledge base. This knowledge base may contain data from user experiences. The objective of the knowledge base is to make the result more accurate and reliable.

1.8 PROFITABLE APPLICATIONS

Data Mining is mainly used today by companies with a strong consumer focus. Some examples are retail, financial, communication, and marketing organizations, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits.

Some areas where data mining is widely used

- **Future Healthcare**

Data mining holds great possibilities to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

- **Market Basket Analysis**

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items, you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer’s needs and change the store’s layout accordingly. Using differential analysis comparison of results between different stores, between customers in different groups can be done.

- **Education**

Educational Data Mining (EDM) concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students’ future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

- **Manufacturing Engineering**

Manufacturing Engineers focus on the design and operation of integrated systems for the production of high-quality, economically competitive products. These systems may include computer networks, robots, machine tools, and materials-handling equipment. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

- **Customer Relationship Management**

Customer Relationship Management (CRM) is about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business needs to collect data and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

- **Fraud Detection**

Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

- **Lie Detection**

Arresting a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This field includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created. With this model processes can be created according to the necessity.

- **Financial Banking**

With computerised banking everywhere, huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

- **Research Analysis**

Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualisation and visual data mining provide us with a clear view of the data.

1.9 DATA MINING TOOLS

Data Mining is the set of techniques that utilize specific algorithms, statical analysis, artificial intelligence, and database systems to analyze data from different dimensions and perspectives

Data Mining tools have the objective of discovering patterns/trends/groupings among large sets of data and transforming data into more advanced information. It is a framework, such as Rstudio or Tableau that allows you to perform different types of data mining analysis. Such a framework is called a data mining tool.

Users can customise data visualisation to suit their business requirement. It features smart drag and drop templates and also includes visualisation maps. Data mining has come a long way, evolving at every step. With various tools already in the market and various other tools which are constantly being added up in the list. Top 10 Free Data Mining Tools for 2021

1. Rapid Miner
2. Orange
3. Weka
4. Sisense
5. Revolution
6. Qlik
7. SAS Data Mining
8. Teradata
9. InetSoft
10. Dundas

1. **Rapid Miner:** So far, this is one of the best tools which uses data to forecast various information. This tool takes up the JAVA language for receiving instructions and is a very insightful tool for predictive analysis. This tool can be used for a lot of functions like training, business applications, etc. This tool makes use of flow-based programming which makes data visualization much easier. It has tools for statistical analysis which are easy to use. Also, one does not need extensive knowledge of coding before using this product.

2. Orange: It is a machine learning software that is component-based and makes data visualization a lot easier. It provides various widgets which analyzes the data and then make it ready for visualization. It has a user engagement platform which is both fun and easy to use. Orange is an open-source data mining platform that can work with both scripts and with ETL workflow. This is one of the simplest tools to operate as it is programmed in Python language which is easier to learn compared to the other programming languages.

3. Weka: This machine learning software is one of the best tools for analyzing data. This tool also aids in predictive modeling and data visualization. This too is written in the JAVA programming language. It can also provide access to various SQL databases which can be analyzed further. It is an open-source tool that is free for use. This tool is mostly used for developing new machine learning algorithms and can support data files from multiple sources. parameters poses a huge challenge.

4. Sisense: It is one of the best artificial intelligence and data aggregation platform. It caters to the needs of different organizations based on the size of the company, the sector in which the company operates, etc. It further combines data from multiple sources and saves it for later use. It also generates visual reports which make the understanding even easier.

5. Revolution: Commonly known as R, this tool provides an interactive platform for statistical operations and data visualization. It is designed in a way that makes it quite user-friendly. It mines the data quite easily. Also, one could perform quite intricate statistical calculations. It makes use of several statistical functions to analyze the data. Also, it makes the heavy programming quite concise and less cumbersome. It has really good graphics features and elements.

6. Qlik: One of the most widely used Business Intelligence tools, Qlik has a easy to use data mining and visualisation platform. This tool allows users to fetch, integrate, process and analyse data from multiple sources.

7. SAS Data Mining: SAS data mining can be used for descriptive and predictive modelling. This tool is especially helpful for developing models quickly, understanding key relations and identifying patterns for streamlining the data mining process. This tool is best suited for text mining and optimisation. It also comes with distributed memory processing architecture that can be scaled up to fit business goals.

8. Teradata: Teradata offers a blend of tools, technologies and expertise that can optimise data mining. Users can integrate this tool into their existing systems and use data from varied sources. This tool is especially beneficial for organisations which have migrated or are migrating to the cloud. Additionally, it supports SQL and provides extensions for data tables. Users can even distribute the data to the discs without much manual intervention.

9. InetSoft: Inetsoft's data mining tool helps users to transform data into uniform data points to aid the analysing process even if it has been sourced variously. This tool can quickly transform data from both structured and unstructured sources . Users can optimise their own apps for updating and data consumption through Inetsoft's on-premise applications. It also allows users to share paginated reports with parameterisation and the corresponding business logic.

10. Dundas: If you are looking for an enterprise-level data mining tool, Dundas could be the perfect fit for you. Dundas can be used for building interactive dashboards, reports and more which can be used at scale. Companies often use it as the central data portal which all the employees can access.

PRACTICE QUESTIONS

1. What you mean by data mining?
2. Explain an example of data mining.
3. What is data mining and how it works?
4. What is data mining and its types?

DATA MINING AND VISUALIZATION

UNIT-II DATA PREPROCESSING

STRUCTURE

2.0 OBJECTIVE

To understand data preprocessing techiuqs and their effect on data

2.1 INTRODUCTION

Data preprocessing is crucial in any data mining process as they directly impact success rate of the project. Data is said to be unclean if it is missing attribute, attribute values, contain noise or outliers and duplicate or wrong data. Presence of any of these will degrade quality of the results. Preprocessing in Data Mining is a data mining technique which is used to transform the raw data in a useful and efficient format. Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Data preprocessing is needed as real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Preprocessing of data is mainly to check the data quality. The quality can be checked by the following

- **Accuracy**: To check whether the data entered is correct or not.
- **Completeness**: To check whether the data is available or not recorded.
- **Consistency:** To check whether the same data is kept in all the places that do or do not match.
- **Timeliness**: The data should be updated correctly.
- **Believability**: The data should be trustable.

Major Tasks in Data Preprocessing:

1. **Data cleaning**
2. **Data integration**
3. **Data reduction**
4. **Data transformation**

Example

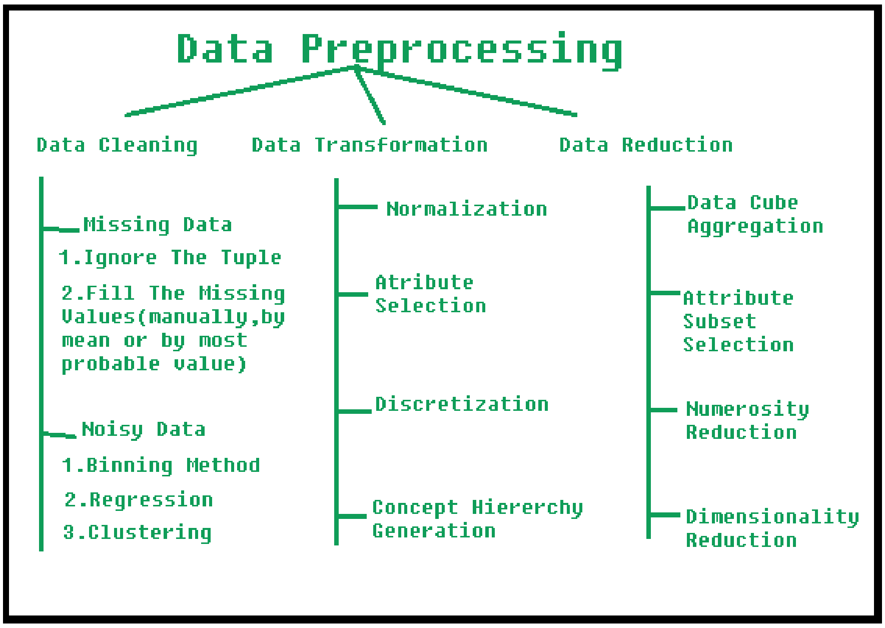
In this example we have 5 Adults in our dataset who have the Sex of Male or Female and whether they are pregnant or not. We can detect that Adult 3 and 5 are impossible data combinations.

		Sex	**Pregnant**
Adult	**1**	Male	No
	2	Female	Yes
	3	**Male**	**Yes**
	4	Female	No
	5	**Male**	**Yes**

We can perform a Data cleansing and choose to delete such data from our table. We remove such data because we can determine that such data existing in the dataset is caused by user entry errors or data corruption. A reason that one might have to delete such data is because the impossible data will affect the calculation or data manipulation process in the later steps of the

		Sex	**Pregnant**
Adult	**1**	Male	No
	2	Female	Yes
	4	Female	No

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:

2.1 DATA CLEANING

The data can have many irrelevant and missing parts. To handle this, data cleaning is done, which 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, which means having one independent variable, or 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.

2.2. DATA TRANSFORMATION

This is needed 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".

2.3 DATA REDUCTION

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 order to get free of this, data reduction technique is used. 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.

This technique is used to aggregate data in a simpler form. For example, imagine that information that is gathered for the 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. Therefore, it can be summarized that that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.

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

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

3. Initial reduced attribute set: { }

4. Step-1: {X1}

5. Step-2: {X1, X2}

6. Step-3: {X1, X2, X5}

7. 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 Forward and Backward Selection

It allows us to remove the worst and select best attributes, saving time and making the process faster.

2 Data Compression

The data compression technique reduces the size of the files using different encoding mechanisms. It can be divided it into two types based on their compression techniques.

- **Lossless Compression**
  Encoding techniques (Run Length Encoding) allows 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 Discrete Wavelet transform technique, PCA (principal component analysis) are examples of this compression. For e.g., JPEG imageoriginal the image. In lossy-data compression, the decompressed data may differ to the original data but are useful enough to retrieve information from them.

3. **Numerosity Reduction**

In this reduction technique the actual data is replaced with mathematical models or smaller representation of the data instead of actual data, it is important to only store the model parameter. Or non-parametric method such as clustering, histogram, sampling. For More Information on Numerosity Reduction Visit the link below:

4. **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 of 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 neighbourhood values in the interval, that process is called bottom-up discretization.

5. **Concept Hierarchies**
   It reduces the data size by collecting and then replacing the low-level concepts (such as 43 for age) to high-level concepts (categorical variables such as middle age or Senior).

For numeric data following techniques can be followed:

6. **Binning**
   Binning is the process of changing numerical variables into categorical counterparts. The number of categorical counterparts depends on the number of bins specified by the user.

7. **Histogram analysis**

Like the process of binning, the histogram is used to partition the value for the attribute X, into disjoint ranges called brackets. There are several partitioning rules:

1. **Equal Frequency partitioning:** Partitioning the values based on their number of occurrences in the data set.
2. **Equal Width Partioning:** Partioning the values in a fixed gap based on the number of bins i.e. a set of values ranging from 0-20.
3. **Clustering:** Grouping the similar data together.

8. **Attribute Subset Selection**

Only 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.

9. **Numerosity Reduction**

This enables to store the model of data instead of whole data, for example: Regression Models.

10. Dimensionality Reduction

This reduces 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)

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 best of the original attributes on the set based on their relevance to other attributes. We know it as a p-value in statistics

2.5 DATA DISCRETIZATION AND CONCEPT HIERARCHY GENERATION

A concept hierarchy for a given numeric attribute attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data y collecting and replacing low-level concepts (such as numeric value for the attribute age) by higher level concepts (such as young, middle-aged, or senior). Although detail is lost by such generalization, it becomes meaningful and it is easier to interpret.

Manual definition of concept hierarchies can be tedious and time-consuming task for the user or domain expert. Fortunately, many hierarchies are implicit within the database schema and can be defined at schema definition level. Concept hierarchies often can be generated automatically or dynamically refined based on statistical analysis of the data distribution. Discretization and Concept Hierarchy Generation for Numeric Data: It is difficult and laborious for to specify concept hierarchies for numeric attributes due to the wide diversity of possible data ranges and the frequent updates if data values. Manual specification also could be arbitrary.

Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis. Five methods for concept hierarchy generation are defined below-binning histogram analysis entropy-based discretization and data segmentation by “natural partitioning

Binning: Attribute values can be discretized by distributing the values into bin and replacing each bin by the mean bin value or bin median value. These technique can be applied recursively to the resulting partitions in order to generate concept hierarchies.

Cluster Analysis: A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms a node of a concept hierarchy, where all noses are at the same conceptual level. Each cluster may be further decomposed into sub-clusters, forming a lower kevel in the hierarchy. Clusters may also be grouped together to form a higher-level concept hierarchy.

Cluster Analysis: A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms a node of a concept hierarchy, where all noses are at the same conceptual level. Each cluster may be further decomposed into sub-clusters, forming a lower kevel in the hierarchy. Clusters may also be grouped together to form a higher-level concept hierarchy.

Data Discretization techniques can be used to divide the range of continuous attribute into intervals. Numerous continuous attribute values are replaced by small interval labels. This leads to a concise, easy-to-use, knowledge-level representation of mining results.

Top-down discretization

If the process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals, then it is called top-down discretization or splitting.

Bottom-up discretization

If the process starts by considering all of the continuous values as potential split-points, removes some by merging neighbourhood values to form intervals, then it is called bottom-up discretization or merging.

To summarize, Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. Another example is analytics, where we gather the static data of website visitors. ...

2.6 CONCEPT HIERARCHIES

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts. In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives.

Data mining on a reduced data set means fewer input/output operations and is more efficient than mining on a larger data set. Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining, rather than during mining.

Transforming nominal data with the use of concept hierarchies allows higher-level knowledge patterns to be found. It allows mining at multiple levels of abstraction, which is a common requirement for data mining applications.

Many concept hierarchies are implicit within the database schema. For example, suppose that the dimension *location* is described by the attributes*number, street, city, province_or_state, zip_ code*, and *country*. These attributes are related by a total order, forming a concept hierarchy such as*"street < city < province_or_state < country"* This hierarchy is shown in Figure 2.5 (a). Alternatively, the attributes of a dimension may be organized in a partial order, forming a lattice. An example of a partial order for the *time* dimension based on the attributes *day, week, month, quarter*, and *year* is *"day* <{*month < quarter; week*} < *year."* This lattice structure is shown in Figure 2.5(b). A concept hierarchy that is a total or partial order among attributes in a database schema is called a schema hierarchy. Concept hierarchies that are common to many applications may be predefined in the data mining system. should provide users with the flexibility to tailor predefined hierarchies according to their particular needs. For example, users may want to define a fiscal year starting on April 1 or an academic year starting on September 1.

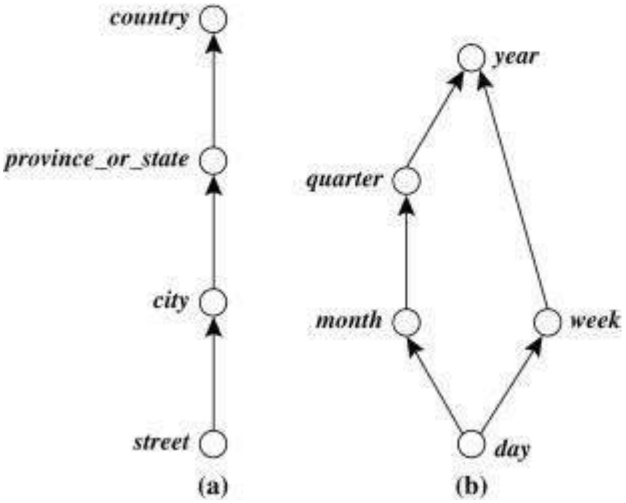


Figure 2.5(a). Figure 2.5(b).

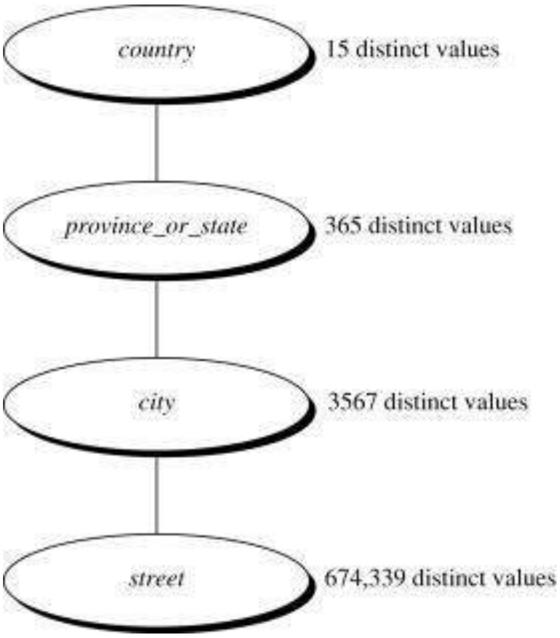
Concept hierarchy generation based on the number of distinct values per attribute

Exmple

Concept hierarchy generation using prespecified semantic connections

Suppose that a data mining expert has held together the five attributes *number, street, city, province_or_state*, and *country*, because they are closely linked semantically regarding the notion of *location*. If a user were to specify only the attribute *city* for a hierarchy defining *location*, the system can automatically drag in all five semantically related attributes to form a hierarchy. The user may choose to drop any of these attributes (e.g. *number* and *street*) from the hierarchy, keeping *city* as the lowest conceptual level.

A concept hierarchy for location can be generated automatically, as illustrated in Figure 2.6. First, sort the attributes in ascending order based on the number of distinct values in each attribute. Second, generate the hierarchy from the top down according to the sorted order, with the first attribute at the top level and the last attribute at the bottom level. Finally, the user can examine the generated hierarchy, and when necessary, modify it to reflect desired attributes. In this example, it is obvious that there is no need to modify the generated hierarchy.



2.7 SUMMARY

In summary, information at the schema level and on attribute–value counts can be used to generate concept hierarchies for nominal data. Transforming nominal data with the use of concept hierarchies allows higher-level knowledge patterns to be found. It allows mining at multiple levels of abstraction, which is a common requirement for data mining applications

The difference between data cleaning and data transformation is that Data cleaning is the process that removes data that does not belong in your dataset. Data transformation is the process of converting data from one format or structure into another. Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format for warehousing and analyzing.

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 map out a framework for your organization.

Step 1: Remove duplicate or irrelevant observations

Step 2: Fix structural errors

Step 3: Filter unwanted outliers

Step 4: Handle missing data

Step 5: Validate and QA

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

- 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 form your next theory?
- If not, is that because of a data quality issue?

2.7 PRACTICE QUESTIONS

1. What is the process of data cleaning?
2. Explain an example of data cleaning.
3. What are the benefits of data cleaning and when data is clean?
4. What is data cleaning and data processing explain with proper example?
5. What is the purpose of concept hierarchy?
6. Why concept hierarchies are useful in data mining?

DATA MINING AND VISUALIZATION

UNIT-III DATA MINING TECHNIQUES

STRUCTURE

3.0 OBJECTIVE

To understand and implement some data mining

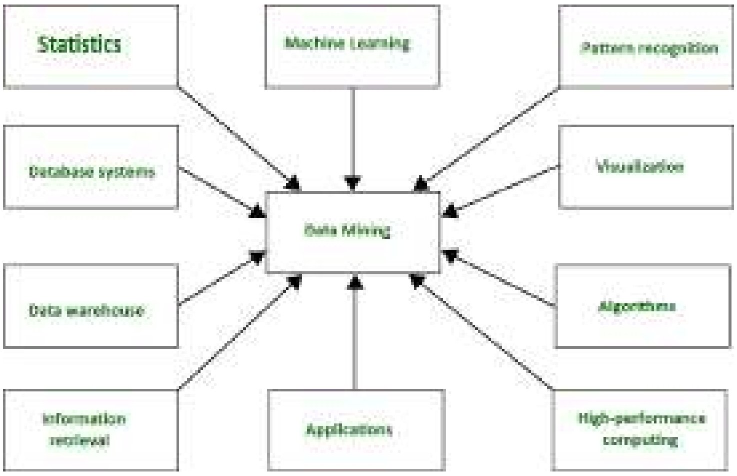
3.1 INTRODUCTION

"**Mining**" is the process of extraction of some valuable material from the earth e.g. coal mining, diamond mining, etc. In the context of computer science, "**Data Mining"** can be referred to as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. It is basically the process carried out for the extraction of useful information from a bulk of data. In the case of coal or diamond mining, the result of the extraction process is coal or diamond. But in the case of Data Mining, the result of the extraction process is not data. Instead, data mining results are the patterns and knowledge that we gain at the end of the extraction process. In that sense, we can think of Data Mining as a step in the process of Knowledge Discovery or Knowledge Extraction.

However, the term 'data mining' became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably.

Nowadays, data mining is used in almost all places where a large amount of data is stored and processed. For example, banks typically use 'data mining' to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well. Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.

Main Purpose of Data Mining



Data Mining

Basically, Data mining has been integrated with many other techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization**,** etc. to gather more information about the data and to helps predict hidden patterns, future trends, and behaviors and allows businesses to make decisions.

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.

Following is a list of Data Mining Techniques: The Complete List

- Data cleaning and preparation.
- Tracking patterns.
- Classification.
- Association.
- Outlier detection.
- Clustering.
- Regression.
- Prediction.

3.2 DATA MINING VERSUS DATABASE MANAGEMENT SYSTEM

Database Management System (DBMS) is a full-fledged system for covering and managing a set of digital databases. It is a collection of computer programs that is dedicated for the management (i.e. organization, storage and retrieval) of all databases that are installed in a system (i.e. hard drive or network).

Data Mining

Data mining is also known as Knowledge Discovery in Data (KDD). It deals with the extraction of previously unknown and interesting information from raw data.

Data Mining

Data mining is the process of analyzing data from a different perspective and summarizing it into useful information – information that can be used to increase revenue cuts cost or both.

Data mining the analysis step of the knowledge discovery in database process. For example, data mining software can help retail companies find customers with a common interest.

The phrase data mining is commonly misused to describe software that presents data in new ways. True data mining software doesn’t just change the presentation, but actually discovers the previously unknown relationship among the data.

Database

The database is a collection of interrelated data and a set of programs to access those data. It is a software system that manages data stored in the database. It provides an effective method of defining, storing and retrieving the information contained in the database. (the primary goal of a DBMS is to provide an environment that is both convenient and efficient to use in retrieving and storing database information. It provides users with information that they required. Some examples of DBMS packages are dBASE, FoxPro, FoxBase, Oracle, Ms-Access etc

Difference

DBMS is a full-fledged system for housing and managing a set of digital databases. However, Data Mining is a technique or a concept in computer science, which deals with extracting useful and previously unknown information from raw data. Most of the times, these raw data are stored in very large databases. Therefore, Data miners use the existing functionalities of DBMS to handle, manage and even preprocess raw data before and during the Data mining process. However, a DBMS system alone cannot be used to analyze data. But, some DBMS at present have inbuilt data analyzing tools or capabilities.

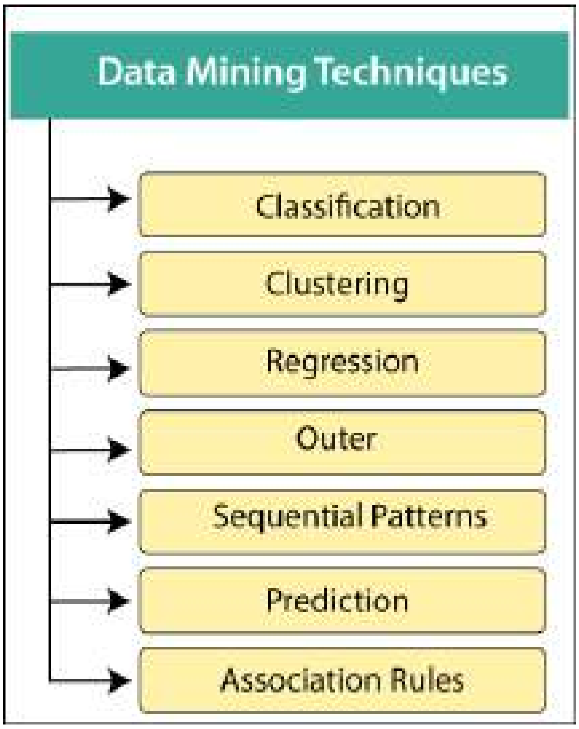
Database	**Data mining**
The database is the organized collection of data. Most of the times, these raw data are stored in very large databases. A Database may contain different levels of abstraction in its architecture. Typically, the three levels: external, conceptual and internal make up the database architecture.	Data mining is analyzing data from different information to discover useful knowledge. Data mining deals with extracting useful and previously unknown information from raw data. The data mining process relies on the data compiled in the data warehousing phase in order to detect meaningful patterns.

Is data mining a database?

Databases are growing in size to a stage where traditional techniques for analysis and visualization of the data are breaking down. Data mining and KDD are concerned with extracting models and patterns of interest from large databases.

3.3 DATA MINING TECHNIQUES

Data mining is a **process of extraction of useful information and patterns from huge data**. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. 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.



3.3.1 Neural Networks

Neural networks are used for effective data mining in order to turn raw data into useful information. Neural networks look for patterns in large batches of data, allowing businesses to learn more about their customers which directs their marketing strategies, increase sales and lowers costs

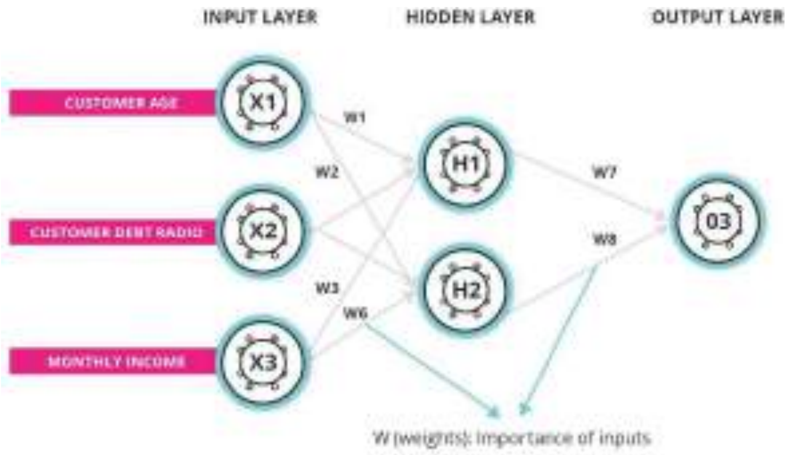


Image source: Mahanta, J. (Jul, 2017). 'Introduction to neural networks, advantages and applications'. Retrieved from Towards Data Science.

It involves a network of simple processing elements that exhibit complex global behavior determined by the connections between the processing elements and element parameters. Artificial neural networks are used with algorithms designed to alter the strength of the connections in the network to produce a desired signal flow.

A neural network is a series of algorithms that attempts to recognize underlying relationships in a set of data through a process that copies the way the human brain operates. Neural networks can adapt to changing input; so that the network generates the best possible result without needing to redesign the output criteria.

Neural Networks in Business

Most retail companies understand that their data provides decision-making information and is a valuable resource in business operations. As technology grows, businesses are leveraging neural networks for predictive analytics to fully harness the benefits of data streams.

Artificial Neural Networks (ANNs) can learn and model non-linear and complex relationships, and have the capacity to manage the reality between the relationship of inputs and outputs, as this is seldom linear or simple. There is a tendency to generalise and attribute unseen relationships on unseen data. ANNs also don’t impose any restrictions on the input variables, unlike other prediction techniques

3.3.2 Association rules

In data science, association rules are used to find correlations and co-occurrences between data sets. They are ideally used to explain patterns in data from seemingly independent information sources, such as relational databases and transactional databases

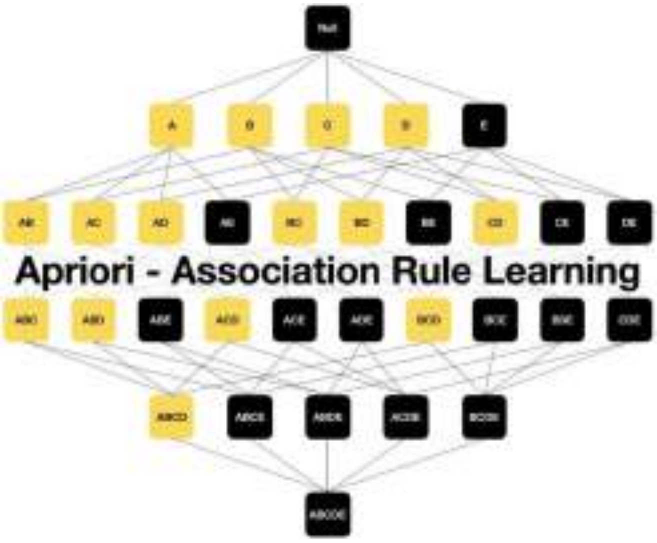
Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

A classic example of association rule mining refers to **a relationship between diapers and beers**. The example, which seems to be fictional, claims that men who go to a store to buy diapers are also likely to buy beer. Data that would point to that might look like this: A supermarket has 200,000 customer transactions. Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

Many algorithms for generating association rules have been proposed. Some well-known algorithms are **Apriori, Eclat and FP-GrowthClassification**

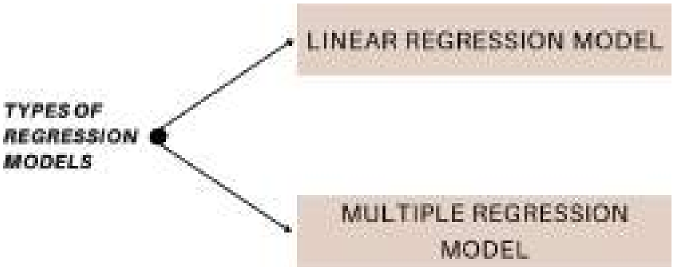
The Apriori algorithm finds association rules in a transaction data of items. A classification dataset, however, is normally in the form of a relational table, which is described by a set of distinct attributes (discrete and continuous)



3.3.3 Regression

Regression is a data mining technique used to predict a range of numeric values (also called continuous values), given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.

Regression techniques consist of finding a mathematical relationship between measurements of two variables, y and x, such that the value of variable y can be predicted from a measurement of the other variable, x



Types of Regression in Data Mining

Two types of Regression can be observed in data mining. Those two types are given below:

1. **Linear Regression Model**
2. **Multiple Regression Model**

Linear Regression is used mainly for the purpose of modeling the relationship between the two given variables. This is usually done by fitting a linear equation to perceive the data.
In addition to that, it can also be used for finding the mathematical relationship between the variables. It is the simplest form of Regression.

Multiple Regression Model

Multiple Regression Model is generally used to explain the relationship between multiple independent or multiple predictor variables.

Applications of Regression

Regression is widely used in many businesses and industries. It is very popular also. Listed below are some of the applications of Regression:
The process of Regression often involves the predictor variable (the values that are identified by the users) and the response variable (the values that are to be predicted).
To summerize, Regression can be defined as a data mining technique that is generally used for the purpose of predicting a range of continuous values (which can also be called "numeric values") in a specific dataset.

3.3.3 Clustering

Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group. Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters. Now that the data from our customer base is divided into clusters, we can make an informed decision about who we think is best suited for this product.

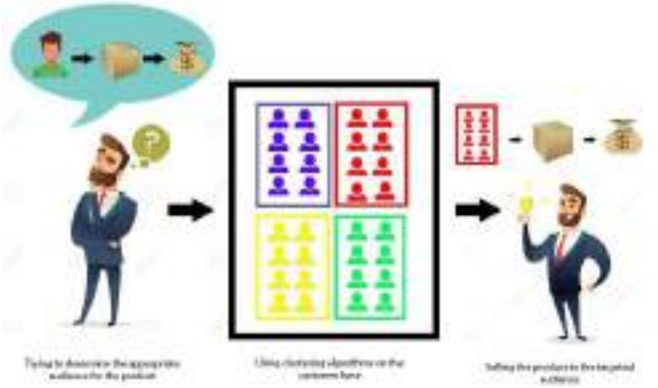
In clustering, a group of different data objects is classified as similar objects. One group means a cluster of data. Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data. The general objective of clustering is to minimize the differences between members of a cluster while also maximizing the differences between clusters.

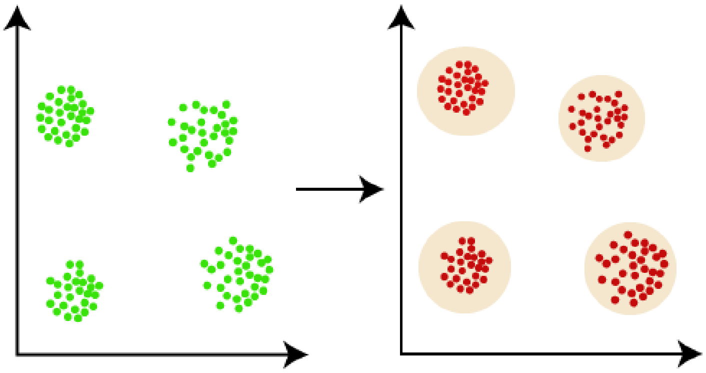
Clustering can help businesses to manage their data better –image segmentation, grouping web pages, market segmentation and information retrieval are four examples. For retail businesses, data clustering helps with customer shopping behavior, sales campaigns and customer retention

Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. Clustering (sometimes called cluster analysis) is usually used to classify data into structures that are more easily understood and manipulated.

Example

Suppose we are a market manager, and we have a new tempting product to sell. We are sure that the product would bring enormous profit, as long as it is sold to the right people. So, how can we tell who is best suited for the product from our company's huge customer base?


Trying to determine the appropriate audience for the product
Using clustering algorithms on the customers base
Selling the product to the targeted audience



Showing four clusters formed from the set of unlabeled data

7 Examples of Clustering Algorithms in Action.

- Identifying Fake News. Fake news is not a new phenomenon, but it is one that is becoming prolific. ...

- Spam filter. ...
- Marketing and Sales. ...
- Classifying network traffic. ...
- Identifying fraudulent or criminal activity. ...
- Document analysis. ...
- Fantasy Football and Sports.

3.4 SUMMARY

- A cluster is a subset of similar objects
- A subset of objects such that the distance between any of the two objects in the cluster is less than the distance between any object in the cluster and any object that is not located inside it.
- A connected region of a multidimensional space with a comparatively high density of objects.
- Clustering is the method of converting a group of abstract objects into classes of similar objects.
- Clustering is a method of partitioning a set of data or objects into a set of significant subclasses called clusters.
- It helps users to understand the structure or natural grouping in a data set and used either as a stand-alone instrument to get a better insight into data distribution or as a pre-processing step for other algorithms

Important points

- ✓ Data objects of a cluster can be considered as one group.
- ✓ We first partition the information set into groups while doing cluster analysis. It is based on data similarities and then assigns the levels to the groups.
- ✓ The over-classification main advantage is that it is adaptable to modifications, and it helps single out important characteristics that differentiate between distinct groups.

3.5 APPLICATIONS OF CLUSTER ANALYSIS IN DATA MINING:

- ➢ data analysis, market research, pattern recognition, and image processing.
- ➢ It assists marketers to find different groups in their client base and based on the purchasing patterns. They can characterize their customer groups.

- It helps in allocating documents on the internet for data discovery.
- Clustering is also used in tracking applications such as detection of credit card fraud.
- It helps in the identification of areas of similar land that are used in an earth observation database and the identification of house groups in a city according to house type, value, and geographical location.

3.6 CLUSTERING IS USED IN DATA MINING FOR THE FOLLOWING REASONS:

Clustering analysis has been an evolving problem in data mining due to its variety of applications. The advent of various data clustering tools in the last few years and their comprehensive use in a broad range of applications, including image processing, computational biology, mobile communication, medicine, and economics, must contribute to the popularity of these algorithms. The main issue with the data clustering algorithms is that it cant be standardized. The advanced algorithm may give the best results with one type of data set, but it may fail or perform poorly with other kinds of data set. Although many efforts have been made to standardize the algorithms that can perform well in all situations, no significant achievement has been achieved so far. Many clustering tools have been proposed so far. However, each algorithm has its advantages or disadvantages and cant work on all real situations.

3.7 QUESTIONS

1. What is clustering in data mining?
2. What are the examples of clustering?
3. What is Cluster Analysis example?
4. What is clustering and its applications?

DATA MINING AND VISUALIZATION

UNIT- IV CLUSTERING

STRUCTURE

4.0 OBJECTIVE

To understand the following clustering techniques

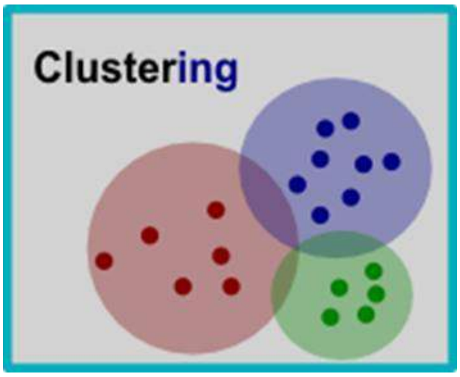
K means, Hierarchical clustering, Agglomerative clustering, Divisive clustering,

4.1 INTRODUCTION

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. Clustering is the process of making a group of abstract objects into classes of similar objects. A cluster of data objects can be treated as one group.

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

It is a main task of exploratory data analysis, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.



4.2 CLUSTER ANALYSIS

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Cluster analysis is the general task to be solved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties

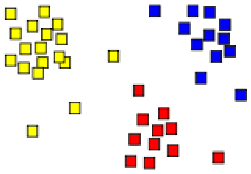


Figure shows the result of a cluster analysis shown as the coloring of the squares into three clusters

Cluster Analysis in Data Mining means that to find out the group of objects which are similar to each other in the group but are different from the object in other groups. While doing cluster analysis, first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

4.3 CLUSTERING METHODS

The method of identifying similar groups of data in a dataset is called clustering. It is one of the most popular techniques in data science. Entities in each group are comparatively more similar to entities of that group than those of the other groups.

Types of clustering algorithms

Since the task of clustering is subjective, this means that can be used for achieving this goal are plenty. Every methodology follows a different set of rules for defining the 'similarit*y;* among data points. In fact, there are more than 100 clustering algorithms known. But few of the algorithms are used popularly:

- **Connectivity models:** As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.
- **Centroid models:** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.
- **Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.
- **Density Models:** These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster.

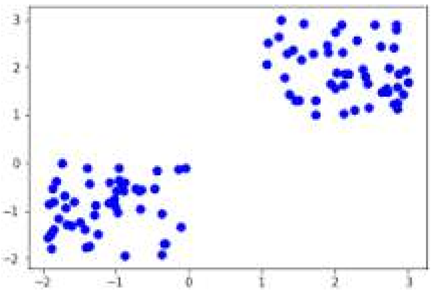
Broadly speaking, clustering can be divided into two subgroups :

- **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not.

- **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

4.3.1 K Means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms......In other words, the K-means algorithm **identifies k number of centroids**, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.



Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster. K-means clustering algorithm computes the centroids and iterates until it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means

K-means clustering is a method of vector quantization, that is popular for cluster analysis in data mining.

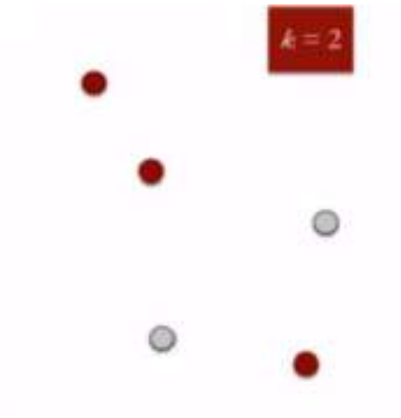
K-means Clustering Method

If k is given, the K-means algorithm can be executed in the following steps:

1. Partition of objects into k non-empty subsets
2. Identifying the cluster centroids (mean point) of the current partition.
3. Assigning each point to a specific cluster
4. Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
5. After re-allotting the points, find the centroid of the new cluster formed.

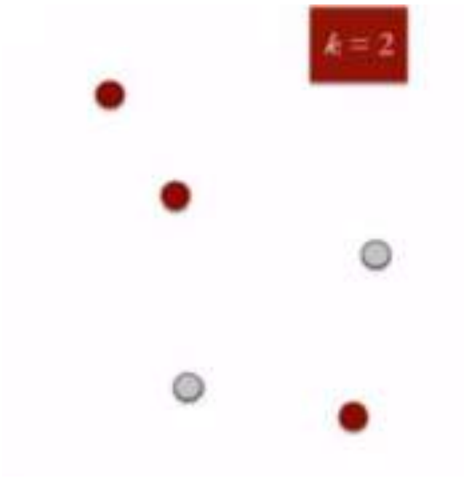
K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps :

1. Specify the desired number of clusters K : Let us choose k=2 for these 5 data points in 2-D space.



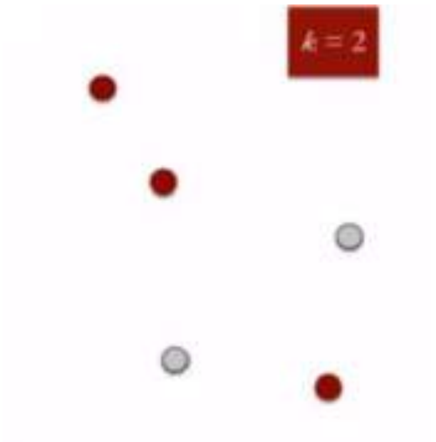
Step 1

2. Randomly assign each data point to a cluster : Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.

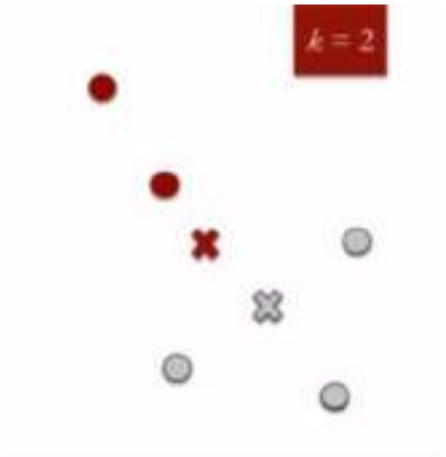


Step 2

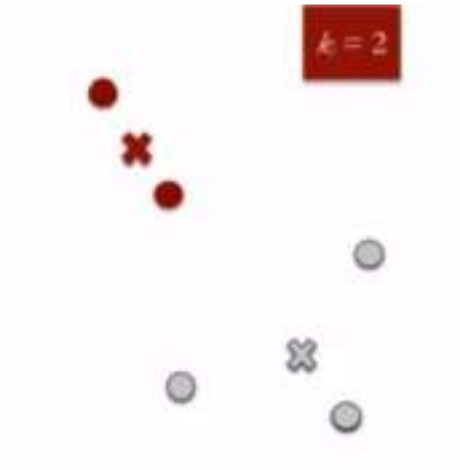
3. Compute cluster centroids : The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



4. Re-assign each point to the closest cluster centroid : Note that only the data point at the bottom is assigned to the red cluster even though its closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster



5. Re-compute cluster centroids : Now, re-computing the centroids for both the clusters.



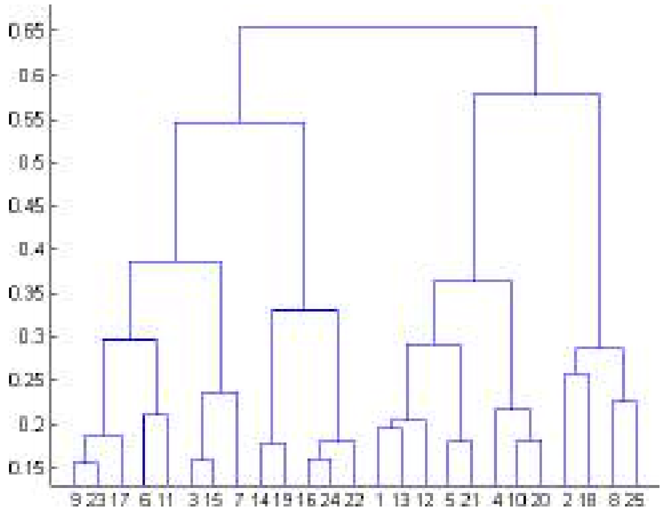
6. Repeat steps 4 and 5 until no improvements are possible : Similarly, repeat the 4th and 5th steps until you reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

4.3.2 Hierarchical Clustering

Hierarchical clustering, as the name suggests is an algorithm that builds hierarchy of clusters. Hierarchical clustering is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

In this algorithm, the hierarchy of clusters is developed in the form of a tree, and this tree-shaped structure is known as the dendrogram. To group the datasets into clusters, it follows the **bottom-up approach**. It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.

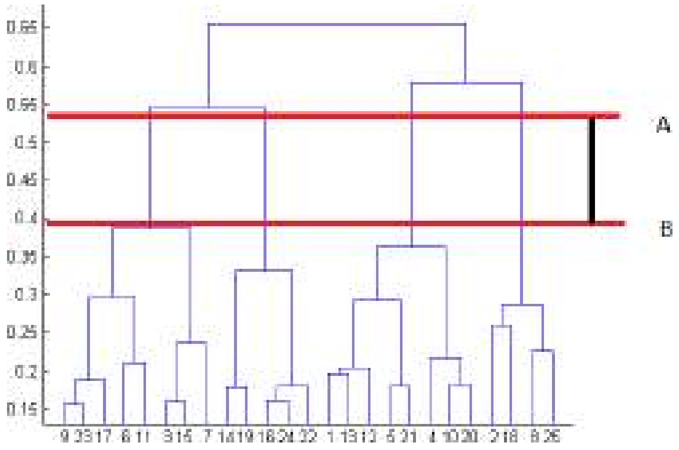
This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. The results of hierarchical clustering can be shown using dendrogram. The dendrogram can be interpreted as:



At the bottom, start with 25 data points, each assigned to separate clusters. Two closest clusters are then merged till we have just one cluster at the top. The height in the dendrogram at which two clusters are merged represents the distance between two clusters in the data space.

The decision of the no. of clusters that can best depict different groups can be chosen by observing the The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

In the above example, the best choice of no. of clusters will be 4 as the red horizontal line in the dendrogram below covers maximum vertical distance AB.



In data mining and statistics, hierarchical clustering analysis is a method of cluster analysis

which seeks to build a hierarchy of clusters i.e. tree type structure based on the hierarchy.

Basically, there are two types of hierarchical cluster analysis strategies

The hierarchical clustering technique has two approaches:

1. **Agglomerative:** Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**

To group the datasets into clusters, it follows the **bottom-up approach**. It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.

4.3.3 Agglomerative Clustering

The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.

The step that Agglomerative Clustering take are:

1. Each data point is assigned as a single cluster.
2. Determine the distance measurement and calculate the distance matrix.
3. Determine the linkage criteria to merge the clusters.
4. Update the distance matrix.
5. Repeat the process until every data point become one cluster.

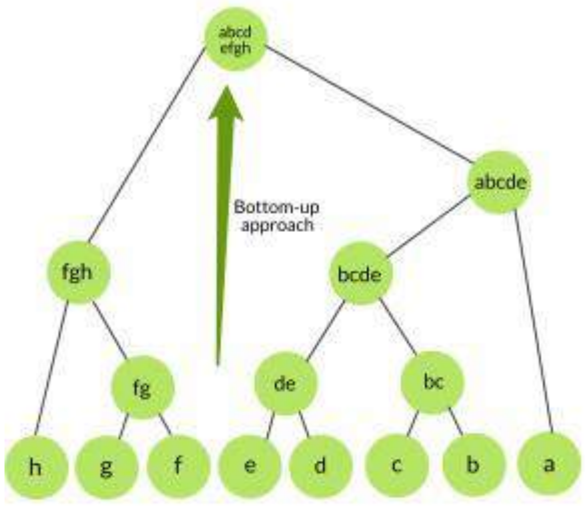
Algorithm for Agglomerative Hierarchical Clustering is: Calculate the **similarity** of one cluster with all the other clusters (calculate proximity matrix) Consider every data point as a individual cluster. Repeat Step 3 and 4 until only a single cluster remains

1. Agglomerative Clustering: Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters.

In data mining and statistics, hierarchical clustering analysis is a method of cluster analysis which seeks to build a hierarchy of clusters i.e. tree type structure based on the hierarchy.

```
for i=1 to N:
  # as the distance matrix is symmetric about
  # the primary diagonal so we compute only lower
  # part of the primary diagonal
  for j=1 to i:
    dis_mat[i][j] = distance[$d_i$, $d_j$]
```

```
each data point is a singleton cluster
repeat
  merge the two cluster having minimum distance
  update the distance matrix
until only a single cluster remains
```



Hierarchical agglomerative clustering

4.3.4 Divisive Clustering

The divisive clustering algorithm is a top-down clustering approach. Initially, all the points in the dataset belong to one cluster and split is performed recursively as one moves down the hierarchy

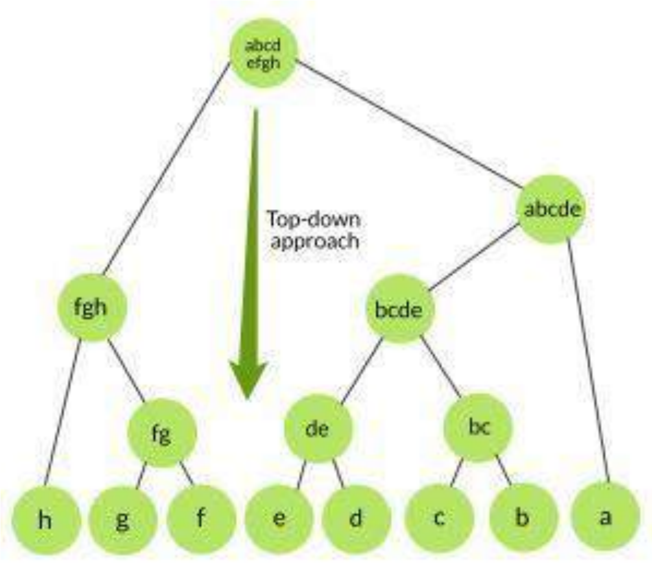
Steps of Divisive Clustering:

1. Initially, all points in the dataset belong to one single cluster.
2. Partition the cluster into two least similar cluster.
3. Proceed recursively to form new clusters until the desired number of clusters is obtained.

2. Divisive clustering: Also known as top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been splitted into singleton cluster.

Algorithm :

```
given a dataset ($d_1$, $d_2$, $d_3$,.....$d_N$) of size N
at the top we have all data in one cluster
the cluster is split using a flat clustering method eg. K-Means etc
repeat
choose the best cluster among all the clusters to split
split that cluster by the flat clustering algorithm
until each data is in its own singleton cluster
```



Hierarchical divisive clustering

Hierarchical Agglomerative *vs* Divisive clustering –

- Divisive clustering is more *complex* as compared to agglomerative clustering, as in case of divisive clustering we need a flat clustering method as "subroutine" to split each cluster until we have each data having its own singleton cluster.
- Divisive clustering is more *efficient* if we do not generate a complete hierarchy all the way down to individual data leaves. Time complexity of a naive agglomerative clustering is **$O(n^3)$** because we exhaustively scan the N x N matrix dist_mat for the lowest distance in each of N-1 iterations. Using priority queue data structure we can reduce this complexity to **$O(n^2 \log n)$**. By using some more optimizations it can be brought down to **$O(n^2)$**. Whereas for divisive clustering given a fixed number of top levels, using an efficient flat algorithm like K-Means, divisive algorithms are linear in the number of patterns and clusters.
- Divisive algorithm is also more *accurate*. Agglomerative clustering makes decisions by considering the local patterns or neighbor points without initially taking into

account the global distribution of data. These early decisions cannot be undone. whereas divisive clustering takes into consideration the global distribution of data when making top-level partitioning decisions.

4.4 EVALUATING CLUSTERS

Cluster Analysis in Data Mining means that to find out the group of objects which are similar to each other in the group but are different from the object in other groups

Here clusters are evaluated based on some similarity or dissimilarity measure such as the distance between cluster points. If the clustering algorithm separates dissimilar observations apart and similar observations together, then it has performed well

Typical objective functions in clustering formalize the goal of attaining high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar). This is an internal criterion for the quality of a clustering.

4.5 SUMMARY

A cluster is a subset of similar objects. A subset of objects such that the distance between any of the two objects in the cluster is less than the distance between any object in the cluster and any object that is not located inside it. A connected region of a multidimensional space with a comparatively high density of objects.

Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group.

Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters. Now that the data from our customer base is divided into clusters, we can make an informed decision about who we think is best suited for this product.

4.6 APPLICATIONS

In many applications, clustering analysis is widely used, such as data analysis, market research, pattern recognition, and image processing. It assists marketers to find different groups in their client base and based on the purchasing patterns. They can characterize their customer groups.

It helps in allocating documents on the internet for data discovery. Clustering is also used in tracking applications such as detection of credit card fraud. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to analyze the characteristics of each cluster.

It helps in the identification of areas of similar land that are used in an earth observation database and the identification of house groups in a city according to house type, value, and geographical location.

4.7 QUESTIONS

1. What is clustering data mining?
2. What are the three types of clusters?
3. How is clustering evaluated?
4. What are the different types of clustering algorithms?
5. What are the major tasks in clustering evaluation?
6. What is clustering assessment?

DATA MINING AND VISUALIZATION

UNIT-V APPLICATIONS OF DATA MINING

STRUCTURE

5.0 OBJECTIVE

To understand some application of data mining techniques

5.1 INTRODUCTION

Data mining is a process of extracting and discovering patterns in large data sets involving methods at the involving machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, dataprerocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures and visualization.

The goal of data mining is the extraction of patterns and knowledge from large amounts of data, not the extraction of data itself.

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining).

Data Mining is mainly used today by companies with a strong consumer focus involving retail, financial, communication, and marketing to handle their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments.

5.2 BUSINESS APPLICATIONS USING DATA MINING

Commercial databases often contain critical business information concerning past performance which could be used to predict the future.

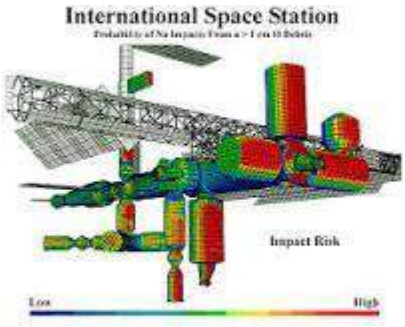
5.2.1 Risk Management and Targeted Marketing,

Risk Management is a logical and systematic method of identifying, analyzing, treating and monitoring the risks involved in any activity or process. The key to successful risk management lies in the ability to modify a formal risk management process that addresses the needs of the systematically addressing risk throughout the product/project life-cycle. Risks can be introduced at the very earliest stages of the project life-cycle. The ability to identify risks earlier translates into earlier risk removal, at less cost, which promotes higher project success probability.

Once risks have been identified and assessed, all techniques to manage the risk fall into one or more of these four major categories:

- Avoidance (eliminate, withdraw from or not become involved)
- Reduction (optimize – mitigate)
- Sharing (transfer – outsource or insure)
- Retention (accept and budget)

Strategies to manage threats (uncertainties with negative consequences) typically include avoiding the threat, reducing the negative effect or probability of the threat, transferring all or part of the threat to another party, and even retaining some or all of the potential or actual consequences of a particular threat. The opposite of these strategies can be used to respond to opportunities (uncertain future states with benefits).


International Space Station
Probability of No Impacts From a > 1 cm ID Debris
Impact Risk
Low
High

Example of risk assessment: A NASA model showing areas at high risk from impact for the International Space Station

5.2.2 Customer Profiles and Feature Construction

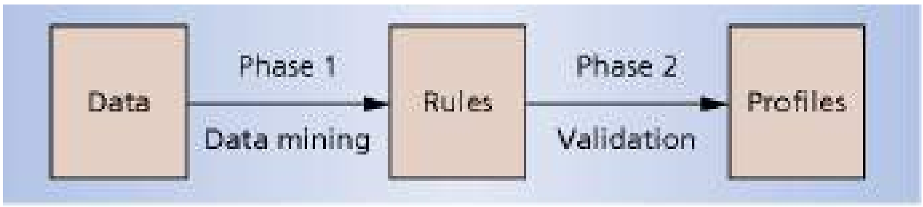
Customer profiling is a method to create a portrait of customers, which includes their personal and transactional details. It helps companies make various customer-centered decisions regarding their business. Customer Profiles or Personas are created with the help of customer research.

In a customer profile, purchasing behaviour of customer is identified, with the intent of targeting similar customers in the sales and marketing campaigns.

However, the huge amounts of data can make the extraction of this business information almost impossible by manual methods or standard software techniques. Data mining techniques can analyze, understand and visualize the huge amounts of stored data gathered from business applications and help stay competitive in today's marketplace.

How Customer Profiling Is Done

Customer profiling is done by breaking customers down into groups that share similar characteristics and goals. For example, if the goal is to purchase a smartphone online, there might not be too much of a difference in the habits of a single professional and a married business executive.



Building Customer profile

Systems construct personal profiles based on customer's transactional histories. System uses data mining techniques to discover a set of rules describing customer's behaviour.

5.2.3 Medical Applications

Data mining it the process of pattern discovery and extraction where huge amount of data is involved. Both the data mining and healthcare industry have emerged some of reliable early detection systems and other various healthcare related systems from the clinical and diagnosis data.

The best procedure for taking data mining beyond the rule of academic research is the three system approach. Implementing all three systems is the way to drive a real-world improvement with any analytics initiative in healthcare. Unfortunately, very few healthcare organizations execute all three of these systems.

These are the following three systems:

The Analytics System:

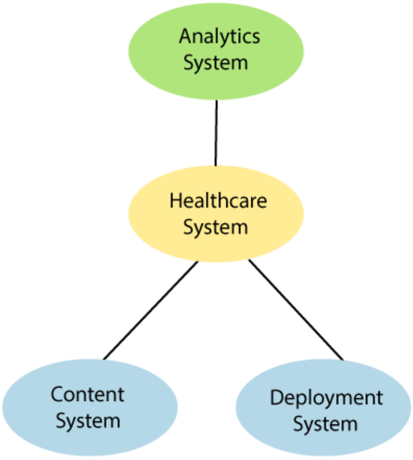
The analytics system incorporates the technology and expertise to accumulate information, comprehend it, and standardize measurements. Aggregating clinical, patient satisfaction, financial, and other data into an enterprise data warehouse (EDW) is the foundation of the system.

The content system:

The content system includes standardizing knowledge work. It applies evidence-based best practices to care delivery. Scientists make significant discoveries each year about clinical best practice, but it mentioned previously, it takes a long time for these discoveries to be incorporated into clinical practice. A strong content system enables organizations to put the latest medical conformation into practice quickly.

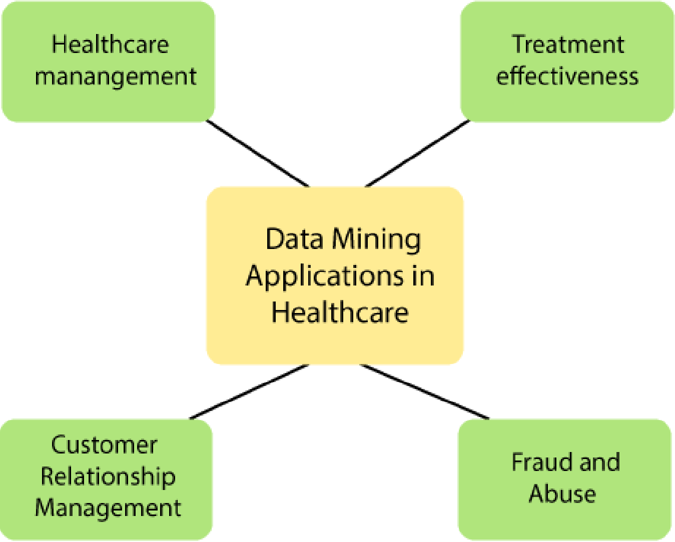
The deployment system:

The deployment system involves driving change management over new hierarchical structures. Particularly, it includes implementing group structures that empower consistently, enterprise-wide deployment of best practices. It requires a real hierarchical change to drive the adoption of best practices throughout an organization.



Application of Data Mining in Healthcare:

Data mining has been used intensively and widely by numerous industries. In healthcare, data mining is becoming more popular nowadays. Data mining applications can incredibly benefit all parties who are involved in the healthcare industry. For example, data mining can help the healthcare industry in fraud detection and abuse, customer relationship management, effective patient care, and best practices, affordable healthcare services. The large amounts of data generated by healthcare transactions are too complex and huge to be processed and analyzed by conventional methods.Data mining provides the framework and techniques to transform these data into useful information for data-driven decision purposes.


Healthcare manangement
Treatment effectiveness
Data Mining Applications in Healthcare
Customer Relationship Management
Fraud and Abuse

Treatment Efectiveness:

Data Mining applications can be used to assess the effectiveness of medical treatments. Data mining can convey analysis of which course of action demonstrates effective by comparing and differentiating causes, symptoms, and courses of treatments.

Healthcare Management:

Data mining applications can be used to identify and track chronic illness states and incentive care unit patients, decrease the number of hospital admissions, and supports healthcare management. Data mining used to analyze massive data sets and statistics to search for patterns that may demonstrate an assault by bio-terrorists.

Customer relationship management:

Customer and management interactions are very crucial for any organization to achieve business goals. Customer relationship management is the primary approach to managing interactions between commercial organizations normally retail sectors and banks, with their customers. Similarly, it is important in the healthcare context. Customer interactions may happen through call centers, billing departments, and ambulatory care settings.

Fraud and abuse:

Data mining fraud and abuse applications can focus on inappropriate or wrong prescriptions and fraud insurance and medical claims.

Results of comparative analysis of various disease in Healthcare:

A comparative analysis of data mining applications in the healthcare sector by various specialists has given in detail. Primarily data mining tools are used to predict the results from the information recorded on healthcare problems. Various data mining tools are utilized to predict the precision level in different healthcare problems. In the given list of medical problems have been examined and evaluated.

Diabetic screening

Diabetes is a major health problem nowadays. Medical applications (diabetic screening), shows that decision tree analyses can be applied to screen individuals for early diabetes risk without the need for offensive tests. This procedure will be particularly useful in developing regions with high epidemiological risk and poor socioeconomic status, and enable clinical practitioners to rapidly screen patients for increased risk of diabetes. The key features in the tree structure could further facilitate diabetes prevention through targeted community interventions, which can improve early diabetes diagnosis and reduce burdens on the healthcare system.

In the present investigation, regression based data mining techniques are applied to diabetes data. The goal of this investigation is to use data mining techniques to discover patterns that identify the best mode of treatment for diabetes across different age groups.

This research concentrates upon predictive analysis of diabetic treatment using a regression-based data mining technique. The Oracle Data Miner (ODM) was employed as a software mining tool for predicting modes of treating diabetes. The support vector machine algorithm was used for experimental analysis. The dataset was studied and analyzed to identify effectiveness of different treatment types for different age groups.. Preferential orders of treatment were investigated. It was concluded that drug treatment for patients in the young age group can be delayed to avoid side effects. In contrast, patients in the old age group should be prescribed drug treatment immediately, along with other treatments, because there are no other alternatives available.

5.2.4 Scientific Applications using Data Mining

An application that simulates real-world activities using mathematics, scientific applications turn real-world objects into mathematical models, and their actions are simulated by executing the required formulas. For example, an airplane's flight characteristics can be simulated in the computer.

5.2.5 Other Applications

Data Mining is primarily used today by companies with a strong consumer focus — retail, financial, communication, and marketing organizations, to bring into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments.

Here is the list of 14 other important areas where data mining is widely used:

1. **Future Healthcare**

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

2. **Market Basket Analysis**

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer’s needs and change the store’s layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

3. Education

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students’ future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

4. Manufacturing Engineering

Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

5. CRM

Customer Relationship Management is all about acquiring and retaining customers, also improving customers’ loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

6. Fraud Detection

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

7. Intrusion Detection

Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity. Data mining also helps extract data which is more relevant to the problem.

8. Lie Detection

Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This filed includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created. With this model processes can be created according to the necessity.

9. Customer Segmentation

Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers. Market is always about retaining the customers. Data mining allows to find a segment of customers based on vulnerability and the business could offer them with special offers and enhance satisfaction.

10. Financial Banking

With computerised banking everywhere huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting,targeting, acquiring, retaining and maintaining a profitable customer.

11. Corporate Surveillance

Corporate surveillance is the monitoring of a person or group's behaviour by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

12. Research Analysis

History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualisation and visual data mining provide us with a clear view of the data.

13. Criminal Investigation

Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

14. Bio Informatics

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

5.3 SUMMARY

Data mining is the process of pattern discovery and extraction where huge amount of data is involved. Both the data mining and healthcare industry have emerged some of reliable early detection systems and other various healthcare related systems from the clinical and diagnosis data.

Various sectors effectively use data mining. It enables the retail sectors to display customer response and helps the banking sector to predict customer profitability. It serves many similar sectors such as manufacturing, telecom, healthcare, automotive industry, education, and many more. Data mining holds incredible potential for healthcare services due to the exponential growth in the number of electronic health records. Previously Doctors and physicians hold patient information in the paper where the data was quite difficult to hold. Digitalization and innovation of new techniques reduce human efforts and make data easily assessable. For example, the computer keeps a massive amount of patient data with accuracy, and it improves the quality of the whole data management system. Still, the major challenge is what should healthcare services providers do to filter all the data efficiently? This is the place where data mining has proven to be extremely useful.

The best procedure for taking data mining beyond the rule of academic research is the three system approach. Implementing all three systems is the way to drive a real-world improvement with any analytics initiative in healthcare. Unfortunately, very few healthcare organizations execute all three of these systems.

5.4 ADVANTAGES OF DATA MINING IN HEALTHCARE:

The data framework simplifies and automates the workflow of health care institutions. Integration of data mining in data frameworks, healthcare institutions reduce decision-making effort and provide new valuable medical knowledge. Predictive models give the best information support and knowledge to healthcare workers. The objective of predictive data mining in medicine is to build up a predictive model that is clear, provides reliable predictions, supports doctors to improve their diagnosis and treatment planning processes. An essential application of data mining is for biomedical signal processing communicated by internal guidelines and reactions to boost the condition, whenever there is a lack of knowledge about the connection between various subsystems, and when the standard analysis methods are ineffective, as it is often in the case of nonlinear associations

5.5 CHALLENGES IN HEALTHCARE DATA MINING:

One of the biggest issues in data mining in healthcare is that the raw medical data is huge and heterogeneous. These data can be accumulated from different sources. For example, from conversations with patients, doctors review, and laboratory results. All these components can have a significant effect on the diagnosis, and treatment of a patient. Missing, incorrect, inconsistent data such as pieces of information saved in various formats from different data sources create a significant obstacle to successful data mining.

Another challenge is that almost all diagnoses and treatments in healthcare are inaccurate and subject to error rates. Here the analysis of specificity and sensitivity are being considered for the measurements of these errors. Within the issue of knowledge integrity evaluation, two major challenges are:

How to create effective algorithms for differentiating the content of two versions (after and before)?

How to create algorithms for evaluating the impact of specific data modifications on the statistical significance of individual patterns, which is collected with the assistance of basic classes of data mining algorithm?

Algorithms that measure the impact that modifications of data values have on the discovered statistical significance of patterns are being created, despite the fact that it is difficult to build up universal measures for all data mining algorithms.

The challenges of Data Mining are as given below:

- Security and Social Challenges.
- Noisy and Incomplete Data.
- Distributed Data.
- Complex Data.
- Performance.
- Scalability and Efficiency of the Algorithms.
- Improvement of Mining Algorithms.

- Incorporation of Background Knowledge.

5.6 PRACTICE QUESTIONS

What is risk taking in marketing management?

What is medical data mining?

What is medical data mining?

What are the data mining applications?

What is the most common application of data mining?

Is data mining used in healthcare?

DATA MINING AND VISUALIZATION

UNIT-VI DATA VISUALIZATION

STRUCTURE

6.0 OBJECTIVE

To understand Data visualization and the rise of HTML 5

6.1 INTRODUCTION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends and patterns in data.

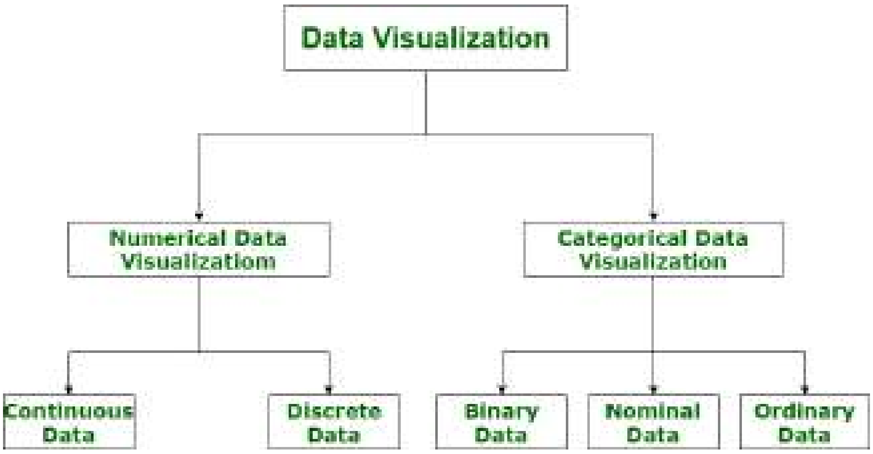
Data visualization is an interdisciplinary field and is a particularly efficient way of communicating when the data is numerous as for example a time series. From an academic point of view, this representation can be considered as a mapping between the original data (usually numerical) and graphic elements (for example, lines or points in a chart). The mapping determines how the attributes of these elements vary according to the data. In this light, a bar chart is a mapping of the length of a bar to a magnitude of a variable. Since the graphic design of the mapping can adversely affect the readability of a chart, mapping is a core competency of Data visualization. Data visualization has its roots in the field of Statistics and is therefore generally considered a branch of Descriptive Statistics. However, because both design skills and statistical and computing skills are required to visualize effectively, it is argued by some authors that it is both an Art and a Science.

Research into how people read and misread various types of visualizations is helping to determine what types and features of visualizations are most understandable and effective in conveying information.Our eyes are attracted to colours and patterns We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that takes our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It's storytelling with a purpose. If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be.

In the "age of Big Data" , visualization is an increasingly key tool to make sense of the trillions of rows of data generated every day. Data visualization helps to tell stories by curating data into a form easier to understand, highlighting the trends. A good visualization tells a story, removing the noise from data and highlighting the useful information. Effective data visualization is a delicate balancing act between form and function. The plainest graph could be too boring to catch any notice or it make tell a powerful point; the most stunning visualization could utterly fail at conveying the right message or it could speak volumes. The data and the visuals need to work together, and there's an art to combining great analysis with

It's hard to think of a professional industry that doesn't benefit from making data more understandable. Every field benefits from understanding data, and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on. While traditional education typically draws a distinct line between creative storytelling and technical analysis, the modern professional world also values those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling.

Due to an increase in statistical data, visual representation of that data is preferred rather than going through spreadsheets. It is easy to understand as well as it saves time, which is of extreme importance. The trends can be easily identified and studied in a graphical representation. Data visualization is mostly used for business analytics. For almost every business, data visualization is used, because they have large data. For the analysis of the data, visuals are the most efficient.



Categories of Data Visualization

6.2 ACQUIRING AND VISUALIZING DATA,

Data viz is the communication of data in a visual manner, or turning raw data into insights that can be easily interpreted by your readers. Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

AN AIR HOSTESS CONNECTS THE GROUP TO A MACHINE WHICH ADMINISTERS A HEAVY SEDATIVE AND PUTS THEM ALL TO SLEEP INSIDE YUSUF (THE CHEMIST'S) DREAM.

The first step in visualizing data is to load it into your application. Typical data sources might be a file on a disk, a stream from a network, or a digitized signal (e.g., audio or sensor readings). The need of immense data is because of various reasons like processing of weeks of data related to surveillance video, one quantitatively acquire data from an hour-long meeting that involved a verbal discussion, drawings on a whiteboard, and note taking done by individual participants?

Thus, the acquisition stage covers several tasks:

The seven stages of visualizing data

Step 1: Define a clear purpose.

Step 2: Know your audience.

Step 3: Keep visualizations simple.

Step 4: Choose the right visual.

Step 5: Make sure your visualizations are inclusive.

Step 6: Provide context.

Step 7: Make it actionable.

There are two kinds Visualization:

1. Interactive Visualization

These days when technological inventions are affecting every market segment, regardless of big or small, interactive visualization needs to be controlled for different portions of charts and graphs to obtain a more detailed analysis of the information being presented.

2. Intuitive Visualization

Data visualization is interactive. Another big picture is available on a click to get a particular information segment. They are also tailored according to the target audience and could be easily updated if the information modifies.

6.3 SIMULTANEOUS ACQUISITION AND VISUALIZATION

Simultaneous acquisition and visualization is needed as millions of visitors take different paths through a website, of a few hundred thousand files on your computer’s hard disk, which ones are taking up the most space, and how often do you use them? By applying methods from the fields of computer science, statistics, data mining, graphic design, and visualization, we can begin to answer these questions in a meaningful way that also makes the answers accessible to others.

The problem is increased by the continually changing nature of data, which can result from new information being added or older information continuously being refined. This flood of data necessitates new software-based tools, and its complexity requires extra consideration.

Each set of data has particular display needs, and the purpose for which it is being used, the data set has effect on those needs as the data itself. There are dozens of quick tools for developing graphics in office programs, on the Web, and elsewhere, but complex data sets used for specialized applications require unique treatment. The characteristics of a data set help determine what kind of visualization is used.

Data acquisition is the **process that serializes the data**, thereby producing data values in a sequence. Visualization is a powerful tool for the qualitative analysis of GCxGC data (e.g., to troubleshoot the chromatography). Various types of visualizations are useful.

Visualization is a powerful tool for qualitative analysis of GC"GC data (e.g., to troubleshoot the chromatography). Various types of visualizations are useful: two-dimensional images provide a comprehensive overview, three-dimensional visualizations effectively illustrate quantitative relationships over a large dynamic range, one-dimensional graphs are useful for overlaying multivariate data, tabular views reveal the numeric values in the data, and graphical and text annotations communicate additional information. This section explores some of the methods and considerations in the various types of visualizations.

Acquire: Obtain the data, whether from a file on a disk or a source over a network.

Parse: Provide some structure for the data's meaning, and order it into categories. The amount of Data one can collect and analyze is immense. It is necessary to put the data you collect it into a structure.This structure will make it easier to know convey to others what data you have by format, tags, names, and indices.

Filter: Remove all but the data of interest. After putting the data into a structure. You will have to filter out the data that is not necessary for your Data Visualization.

Mine: Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context. The focus will be on basic statistics in the beginning. More emphasis in discovering patterns will occur in later sections. This step helps get basic understanding of the data before doing the representational step.

Represent: Choose a basic visual model, such as a bar graph, list, or tree. Focus on how to choose the basic visual model as well as how to represent it.

Refine: Improve the basic representation to make it clearer and more visually engaging.

Interact: Add methods for manipulating the data or controlling what features are visible.

6.4 APPLICATIONS OF DATA VISUALIZATION,

Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. Data Visualization Application enables users to visualize data, draw insights and understand it better. It allows people to organize and present information intuitively. Visualizations can be combined into a dashboard. Dashboards are useful because they allow you to relate different views of information visually.

Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.

- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

Data Mining System Products

There are many data mining system products and domain specific data mining applications. The new data mining systems and applications are being added to the previous systems. Also, efforts are being made to standardize data mining languages.

Choosing a Data Mining System

The selection of a data mining system depends on the following features −

- **Data Types** − The data mining system may handle formatted text, record-based data, and relational data. The data could also be in ASCII text, relational database data or data warehouse data. Therefore, we should check what exact format the data mining system can handle.

- **System Issues** − We must consider the compatibility of a data mining system with different operating systems. One data mining system may run on only one operating system or on several. There are also data mining systems that provide web-based user interfaces and allow XML data as input.
- **Data Sources** − Data sources refer to the data formats in which data mining system will operate. Some data mining system may work only on ASCII text files while others on multiple relational sources. Data mining system should also support ODBC connections or OLE DB for ODBC connections.
- **Data Mining functions and methodologies** − There are some data mining systems that provide only one data mining function such as classification while some provides multiple data mining functions such as concept description, discovery-driven OLAP analysis, association mining, linkage analysis, statistical analysis, classification, prediction, clustering, outlier analysis, similarity search, etc.
- **Coupling data mining with databases or data warehouse systems** − Data mining systems need to be coupled with a database or a data warehouse system. The coupled components are integrated into a uniform information processing environment. Here are the types of coupling listed below
  - No coupling
  - Loose Coupling
  - Semi tight Coupling
  - Tight Coupling
- **Scalability** − There are two scalability issues in data mining −
  - **Row (Database size) Scalability** − A data mining system is considered as row scalable when the number or rows are enlarged 10 times. It takes no more than 10 times to execute a query.

6.5 KEY FACTORS OF DATA VISUALIZATION

6.5.1 Control of Presentation

Visualization is the first step to make sense of data. To translate and present data and data correlations in a simple way, data analysts use a wide range of techniques. Some of them are charts, diagrams, maps, etc. Choosing the right technique and its setup is often the only way to make data understandable.

Five factors that influence data visualization choices:

1. **Audience**
   Adjust data representation to the specific target audience. For example, fitness mobile app users who browse through their progress can easily work with uncomplicated

visualizations. On the other hand, if data visions are intended for researchers and experienced decision-makers who regularly work with data,

2. **Content**
   The type of data to be presented will determine the strategies. For example, if it's time-series metrics, you will use line charts to show the dynamics in many cases. To show the relationship between two elements, scatter plots are often used. In turn, bar charts work well for comparative analysis.
3. **Context**
   You can use different data visualization approaches and read data depending on the context. To emphasize a certain figure, for example, significant profit growth, you can use the shades of one color on the chart and highlight the highest value with the brightest one.
4. **Dynamics**
   There are various types of data, and each type has a different rate of change. For example, financial results can be measured monthly or yearly, while time series and tracking data are changing constantly. Depending on the rate of change, you may consider dynamic representation (steaming) or static visualization techniques in data mining.
5. **Purpose**
   The goal of data visualization affects the way it is implemented. In order to make a complex analysis, visualizations are compiled into dynamic and controllable dashboards that work as visual data analysis techniques and tools. However, dashboards are not necessary to show a single or occasional data insight.

Depending on these factors, different data visualization techniques can be used. Here are the common types of visualization techniques:

- **Charts**
  The easiest way to show the development of one or several data sets is a chart. Charts vary from bar and line charts that show the relationship between elements over time.

- **Plots**
  Plots allow to distribute two or more data sets over a 2D or even 3D space to show the relationship between these sets and the parameters on the plot.

- **Maps**
  They allow to locate elements on relevant objects and areas — geographical maps, building plans, website layouts, etc.

- **Diagrams and matrices**
  Diagrams are usually used to demonstrate complex data relationships and links and include various types of data on one visualization.

Faster and Better JavaScript processing, For a Java Script (JS) developer, the ability to visualize data is just as valuable as making interactive Web pages. Especially that the two often go in pairs. As JavaScript continues to gain popularity in data visualization domain, the market is flooded with even new libraries with which to create beautiful charts for the Web

Data is visualized in JavaScript by using the following method:

Write the Code

1. Build the HTML.
2. Understand the Data.
3. JavaScript to Load the Data.
4. Understand the Algorithm.
5. Build the Data Table with JavaScript.
6. Add the Data to the Table with JavaScript.
7. Add the Color Legend.
8. Style the Visualization with CSS.

The form of the visualization depends on the relationship between the chosen graphical elements or types and the data points involved. The end goal is to communicate information derived from data sources clearly and effectively via visual or graphical means.

Following are some JavaScript's data visualization libraries

- Highcharts
- Toast UI Chart
- D3.js
- Recharts
- Chart.js

6.5.2 Rise of HTML5

Rise HTML is **a Learning Journal** that allows the learner to enter text responses to journal prompts throughout a Rise course and at the end print their learning journal of all their responses. If you are familiar with Javascript, you can build on it. Learning Journal in Rise. Rise 360 is a modern, dynamic eLearning authoring tool allowing designers to create responsive courses for any device. Using a web-based course builder, Rise allows instructional designers to create beautiful online courses with a few clicks of a button.

Rise training is the all-in-one training system that makes training easy to create, enjoyable to take, and simple to manage. Taking a closer look at why Rise is all that is needed to create, distribute, track, and manage online training people actually love.

Meeting Preparation

Heading

When we show up to the present moment with all of our senses, we invite the world to fill us with joy. The pains of the past are behind us. The future has yet to unfold. But the now is full of beauty simply waiting for our attention.

6.6 SUMMARY

The types of data Visualisation are Column Chart. This is one of the most common types of data visualization tools. Bar Graph, Stacked Bar Graph, Line Graph, Dual-Axis Chart, Pie Chart, Scatter Plot.

Data Visualization process is a series of steps:

Acquire, Parse, Filter, Mine, Represent, Refine, Interact

Data mining concepts are still evolving and here are the latest trends that we get to see in this field −Application Exploration, Scalable and interactive data mining methods, Integration of data mining with database systems, data warehouse systems and web database systems, Visual data mining, New methods for mining complex types of data, Web mining, Distributed data mining, Real time data mining.

6.7 Practice Exercise

1. What is data visualization?
2. What are data visualization techniques?
3. Where is data visualization used?
4. What is acquiring data and visualize data?
5. What are the seven stages of visualizing data?

MODULE - VII

EXPLORING THE VISUAL DATA SPECTRUM

:: TOPICS ::

- Data Points
- Line Chart
- Bar Chart
- Pie Chart
- Area Chart
- Candlestick Chart
- Bubble Chart
- Surface Plot
- Map Chart
- Infographics

Dr. Chetan R. Dudhagara
Assistant Professor and Head
Department of Communication & Information Technology
International Agribusiness Management Institute
Anand Agricultural University
Anand, Gujarat, India

DATA MINING AND VISUALIZATION

Unit – VII

EXPLORING THE VISUAL DATA SPECTRUM

<u>Structure</u>

OBJECTIVES

The main objective of this module is to understand the data mining and visualization concepts. The various data visualization techniques and charts such as data points, line chart, bar chart, pie chart and area chart are covered in this module. The advance data visualization techniques are also covered in this module such as candlestick chart, bubble chart, surface chart, map chart and infographics.

1. INTRODUCTION

Data science become a buzzword that everyone talks about the data science. Data science is an interdisciplinary field that combines different domain expertise, computer programming skills, mathematics and statistical knowledge to find or extract the meaningful or unknown patterns from unstructured and structure dataset. Data science is useful for extraction, preparation, analysis and visualization of various information. Various scientific methods can be applied to get insight in data.

Data mining is a process to find the new and hidden patterns or relationship in a large data sets for solving various business or real-life problems to improve the efficiency. The various data mining software or tools are used such as Rapid Miner, SPSS, Oracle, Orange, Weka, R, Python, etc.

Data mining is used in various real-life application such as banking, marketing, retail, health care, agriculture, insurance, transportation etc.

Data visualization is a graphical representation of data and quantitative information by using various graphical tools such as graphs, charts and maps. It is more useful to understand the trends, patterns and outliers in the data.

2. EXPLORING DATA VISULIZATION

Charts is the representation of data in a graphical form. It helps to summarizing and presenting a large amount of data in a simple and easy to understandable formats. By placing the data in a visual context, we can easily detect or identify the patterns, trends and correlations among them.

Python provides various easy to use multiple graphics libraries for data visualization. These libraries are work with both small and large datasets.

Python has multiple graphics libraries with different features. Some of the most popular and commonly used Python data visualization libraries are Matplotlib, Pandas, Seaborn, Plotly and ggplot.

Matplotlib is a most popular, amazing and multi-platform data visualization library available in Python. It consists a wide variety of plots like data points, line chart, bar chart, pie chart, area chart etc.

2.1 Data Points / Scatter Plot

Data points is a mark in a diagram where each value in the dataset is representing by a dot or point. It is a set of dotted points to represent the individual data on both horizontal and vertical axis to reveal the distribution trends of data.

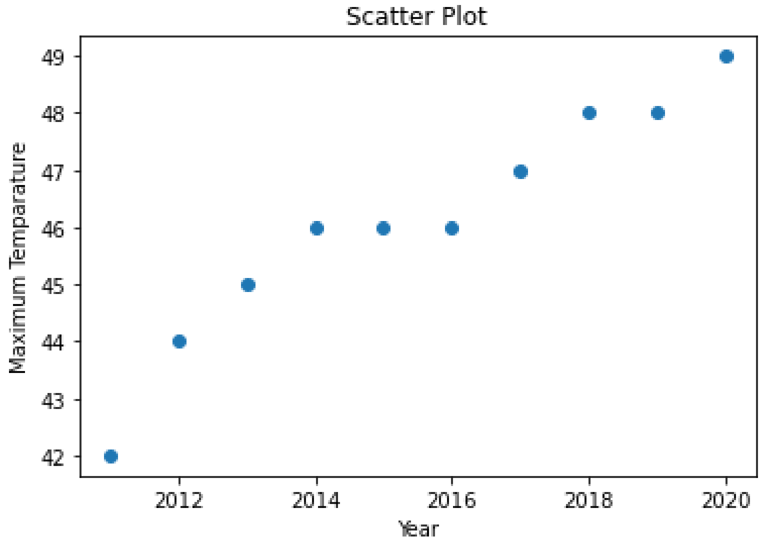
This plot is mostly used for large dataset to highlight the similarities in a dataset. It also shows the outliers and distribution of data.

The ***scatter()*** function is used to draw the scatter plot. This function plots one dot for each observation. It requires two different arrays of same length for both x-axis and y-axis. we can also set the scatter plot title and labels on both the axis.

Example: Here we take an example of maximum temperature noted every year. The x-axis represents **"Year"** values and y-axis represents **"Max_Temp"**. (**Note:** The maximum temperature is a dummy data use for example purpose only)

```
# Importing library
import matplotlib.pyplot as plt

# Data values
Year = [2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,
2019, 2020]
Max_Temp = [42,44,45,46,46,46,47,48,48,49]

# Plotting scatter plot with title and label
plt.scatter(Year, Max_Temp)
plt.title("Scatter Plot")
plt.xlabel("Year")
plt.ylabel("Maximum Temperature")
plt.show()
```

The above code will create scatter plot as follow:

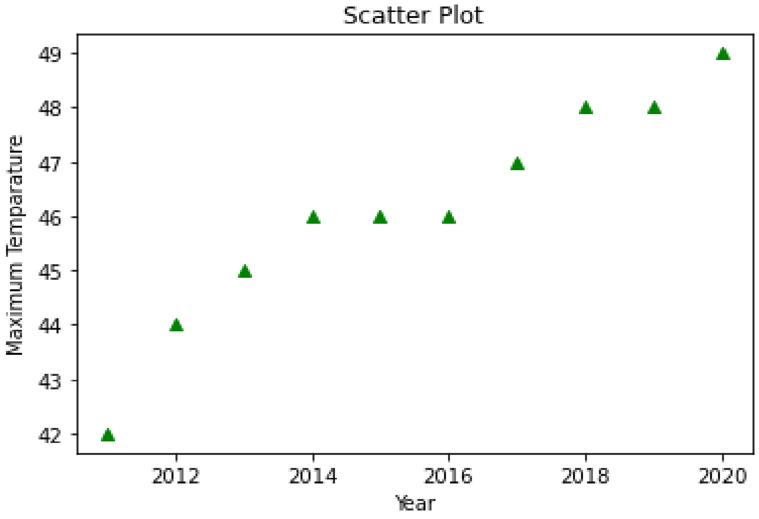


In above plot, we can see the trends of maximum temperature every year.

We can also set or change the color using ***color*** as an argument. Here we set the ***green*** color for temperature. We can also set the shape of data points using ***marker*** as an argument as follows:

```
# Importing library
import matplotlib.pyplot as plt

# Data values
Year = [2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,
2019, 2020]
Max_Temp = [42,44,45,46,46,46,47,48,48,49]

# Plotting scatter plot with title and label
plt.scatter(Year, Max_Temp, color="Green", marker="^")
plt.title("Scatter Plot")
plt.xlabel("Year")
plt.ylabel("Maximum Temperature")
plt.show()
```

The above code will create scatter plot as follow:



2.2 Line Chart

Line chart is used to show the relation between two datasets on a different axis. There are multiple features available such as line color, line style, line width etc.

Matplotlib is a most popular library for plotting different chart. Line chart is one of them.

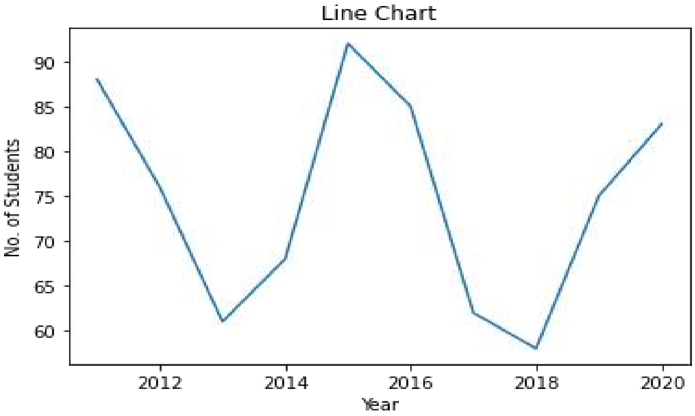
The ***plot()*** function is used to create a line chart in Python. Here we will see some examples of line chart in Python.

Example: Here we take an example of numbers of students enroll in specific course in different year. The x-axis represents **"Year"** values and y-axis represents **"Student"**.

```
# Importing library
from matplotlib import pyplot as plt

# Data values
```

```
Year = [2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,
2019, 2020]
Student = [88,76,61,68,92,85,62,58,75,83]

# Plotting the line chart
plt.title("Line Chart")
plt.xlabel("Year")
plt.ylabel("No. of Students")
plt.plot(Year, Student)
plt.show()
```

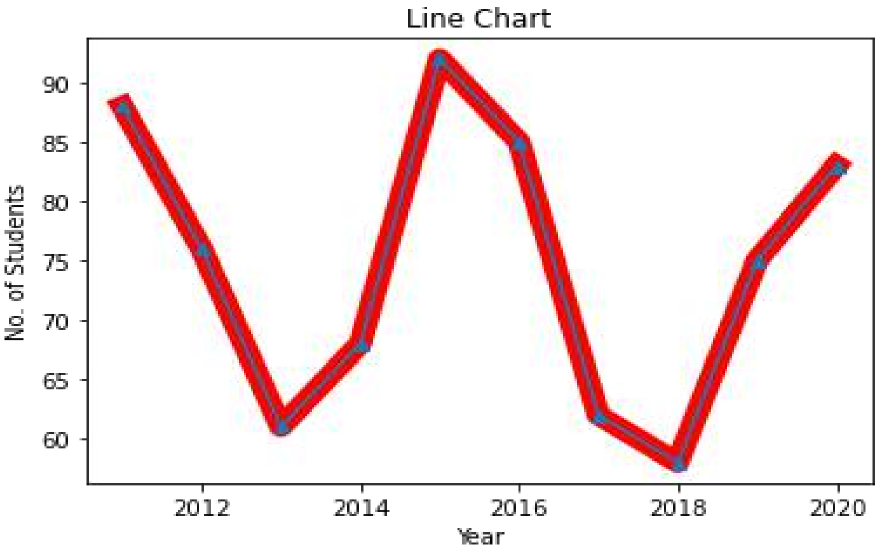
The above code will create line chart as follow:



We can set the line color using ***color*** as an argument. Here we set ***red*** color to the line. We can also set the shape of data points using ***marker*** as an argument. We can also set the line width using ***linewidth*** or ***lw*** as an argument. Here we set ***10*** to the linewidth as follows:

```
# Importing library
from matplotlib import pyplot as plt

# Data values
Year = [2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,
2019, 2020]
Student = [88,76,61,68,92,85,62,58,75,83]

# Plotting the line chart
plt.title("Line Chart")
plt.xlabel("Year")
plt.ylabel("No. of Students")
plt.plot(Year, Student, color="Red", marker="^",
linewidth=10)
plt.show()
```

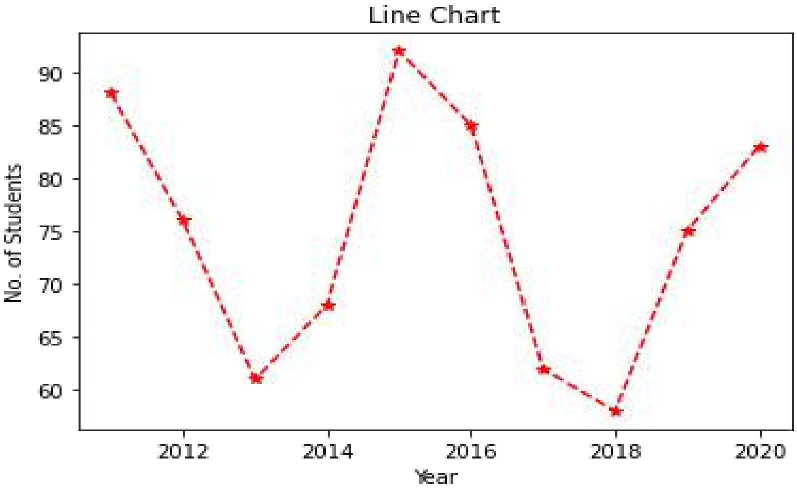
The above code will create line chart as follow:



We can set the line style also using ***linestyle*** or ***ls*** as an argument. There are various types of style available such as solid, dotted, dashed and dashdot. Here we set ***dashed*** as a linestyle in a line chart.

```
# Importing library
from matplotlib import pyplot as plt
# Data values
Year = [2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018,
2019, 2020]
Student = [88,76,61,68,92,85,62,58,75,83]
# Plotting the line chart
plt.title("Line Chart")
plt.xlabel("Year")
plt.ylabel("No. of Students")
plt.plot(Year, Student, marker="*", linestyle =
'dashed', color="Red")
plt.show()
```

The above code will create line chart as follow:



2.3 Bar Chart

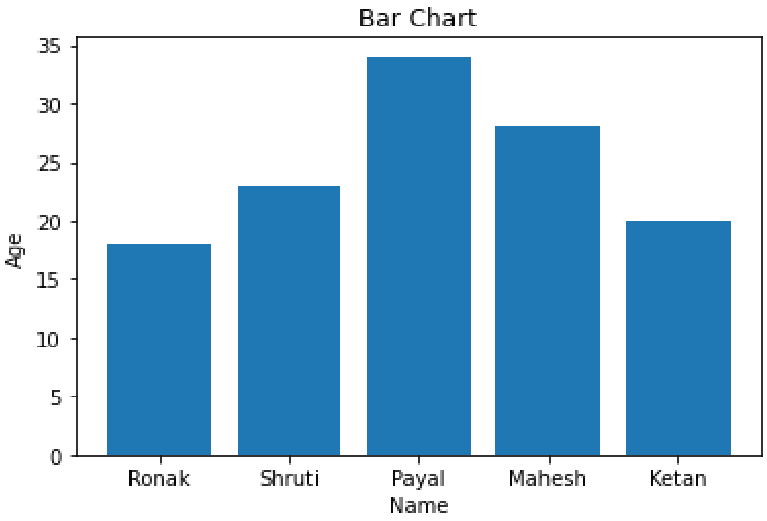
Bar chart or bar plot is representing the category of data with rectangular bars with different heights and lengths with reference to the values of that they present. The ***bar()*** function is used to create a bar chart. The bar chart can be plotted both horizontally and vertically.

The bar chart describes the comparisons between distinct categories. One axis represents the particular categories being compared and another axis represent the measured values respected to those categories. The numerical values of variables in a dataset represent the height or length of bar.

Example: Here we take an example of students name and age. The x-axis represents **"Name"** and y-axis represents **"Age"**. Here we also set the chart title and labels on both the axis.

```
# Importing library
from matplotlib import pyplot as plt

# Data values
Name = ["Ronak", "Shruti", "Payal", "Mahesh", "Ketan"]
Age = [18, 23, 34, 28, 20]

# Labelling the axes and title
plt.title("Bar Chart")
plt.xlabel("Name")
plt.ylabel("Age")

# Plotting the bar chart
plt.bar(Name, Age)
```

The above code will create bar chart as follow:

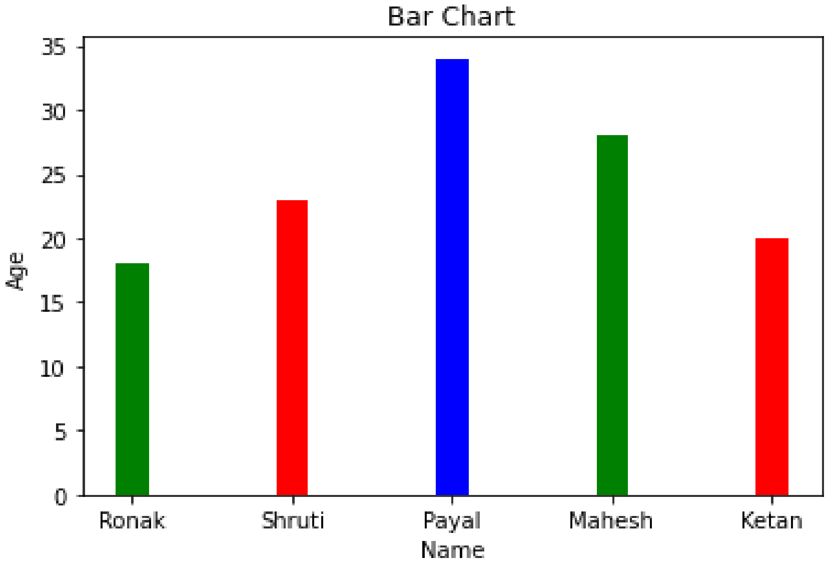


We can change the bar color also. We set ***color*** as an argument to change the color of bar. Here we set the multiple color of bar. We can set the bar width also. We set ***width*** as an argument to set the width of bar. Here we set ***0.2*** as a width of bar.

```
# Importing library
from matplotlib import pyplot as plt
```

```
# Data values
Name = ["Ronak", "Shruti", "Payal", "Mahesh", "Ketan"]
Age = [18, 23, 34, 28, 20]

# Labelling the axes and title
plt.title("Bar Chart")
plt.xlabel("Name")
plt.ylabel("Age")

# Plotting the bar chart
plt.bar(Name, Age, width=0.2, color = ["Green", "Red",
"Blue"])
```

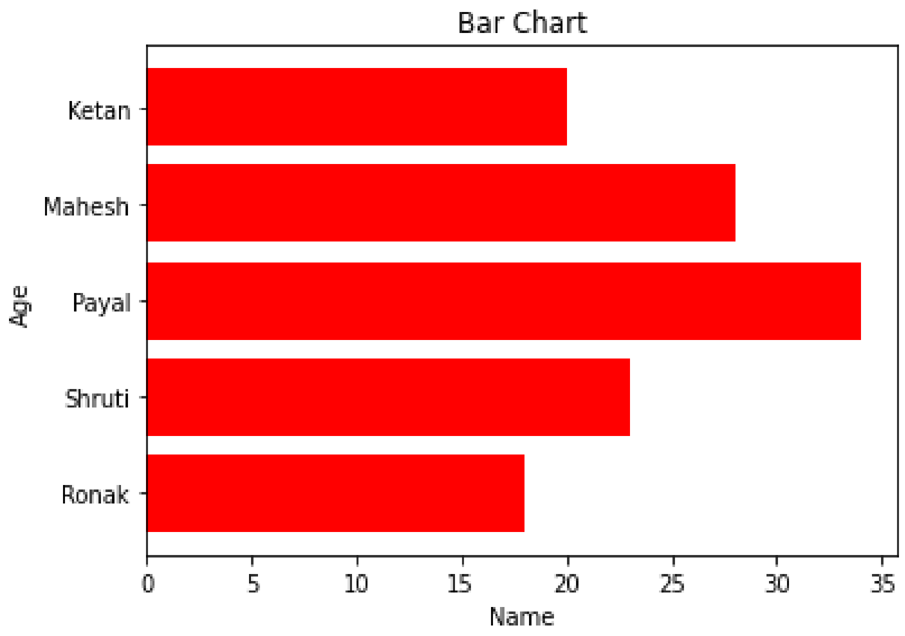
The above code will create bar chart as follow:



We can display bar horizontally also instead of vertically. We set the bar horizontally by using ***barh()*** function.

```
# Importing library
from matplotlib import pyplot as plt

# Data values
Name = ["Ronak", "Shruti", "Payal", "Mahesh", "Ketan"]
Age = [18, 23, 34, 28, 20]

# Labelling the axes and title
plt.title("Bar Chart")
plt.xlabel("Name")
plt.ylabel("Age")

# Plotting the bar chart
plt.barh(Name, Age, color="Red")
```

The above code will create bar chart as follow:



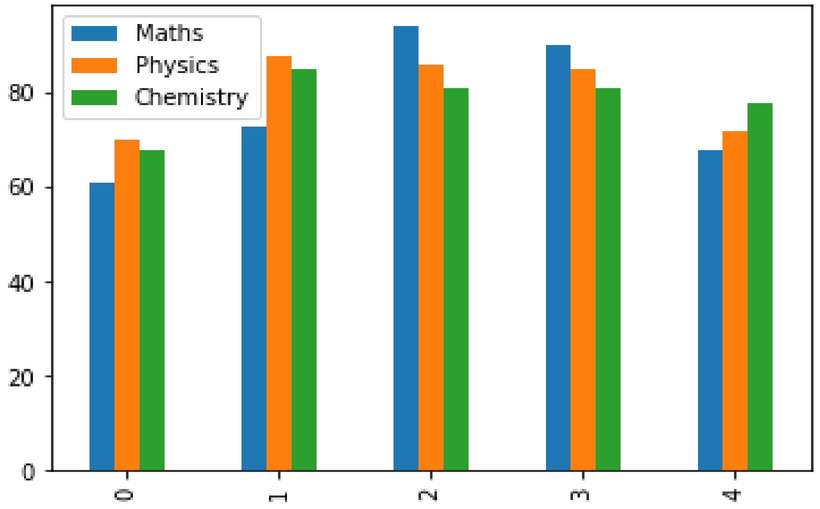
The multiple bar chart is used to represent the comparison among the different variables in a dataset.

Here, we take an example of marks of different subject with the name of students. The x-axis represents students and y-axis represents marks of different subjects.

```
# Importing library
from matplotlib import pyplot as plt
import pandas as pd

# Data values
df = pd.DataFrame({
"Name":["Ronak", "Shruti", "Payal", "Mahesh", "Ketan"],
"Maths":[61,73,94,90,68],
"Physics":[70,88,86,85,72],
"Chemistry":[68,85,81,81,78]})

# Plotting the bar chart
df.plot.bar()
```

The above code will create bar chart as follow:

Here, we show the comparisons of marks of different subjects of the students.



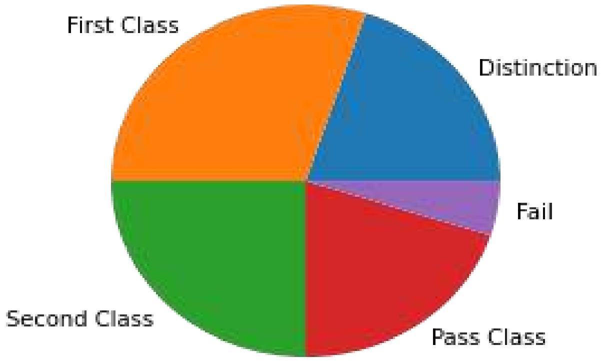
2.4 Pie Chart

The pie chart is a circular analytical or statistical type of graphs that represents the data in a circular plot. The total area in a pie chart is divided into different region or slices to symbolize the numerical percentage. The slices of pie are called wedges.

Example: Here we take an example of results of the students. The results of students are in different categories such as distinction, first class, second class, pass class and fail. Here area represent the percentage value of results of the students respectively.

```
# Importing library
import matplotlib.pyplot as plt

# Labels and area covered by each label
result = ["Distinction", "First Class", "Second Class",
"Pass Class", "Fail"]
area = [20,30,25,20,5]

# Plotting pie chart
plt.pie(area, labels=result)
plt.show()
```

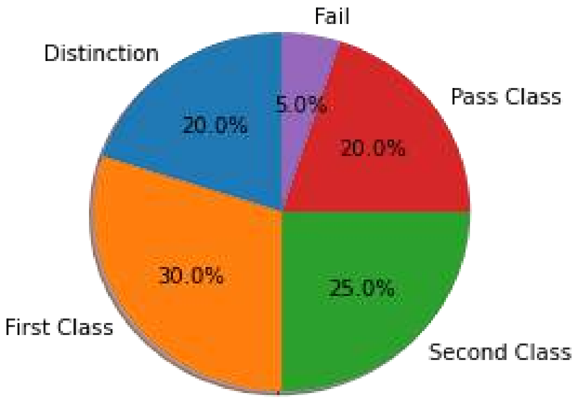
The above code will create pie chart as follow:



We can set the ***startangle*** argument to start the rotation of pie chart by given degrees. Here we set the ***startangle*** is ***90***. We can also set the ***autopct*** argument to show the percentage value. We can also set the ***shadow*** of pie chart.

```
# Importing library
import matplotlib.pyplot as plt

# Labels and area covered by each label
result = ["Distinction", "First Class", "Second Class",
"Pass Class", "Fail"]
area = [20,30,25,20,5]

# Plotting pie chart
plt.pie(area, labels=result, startangle=90,
autopct='%1.1f%%', shadow=True )
plt.show()
```

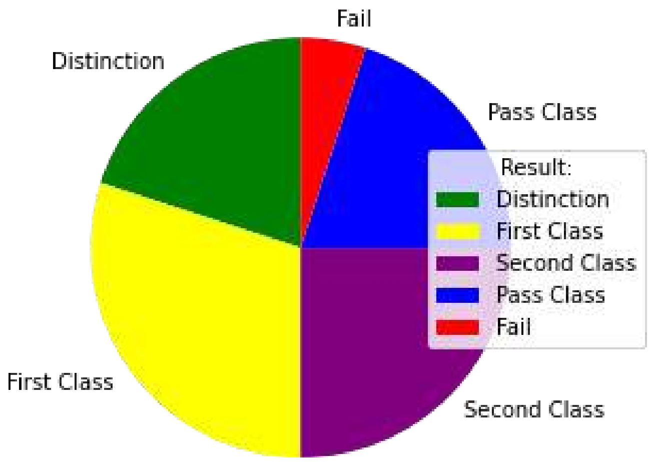
The above code will create pie chart as follow:



We can set the different colors of each slices in a pie chart using ***colors*** argument. We can also set the legend and location of legend in pie chart.

```
# Importing library
import matplotlib.pyplot as plt

# Labels and area covered by each label
result = ["Distinction", "First Class", "Second Class",
"Pass Class", "Fail"]
area = [20,30,25,20,5]
mycolor = ["Green", "Yellow", "Purple", "Blue", "Red"]

# Plotting pie chart
plt.pie(area, labels=result, startangle=90,
colors=mycolor)
plt.legend(title="Result:", loc="right")
plt.axis('equal')
plt.show()
```

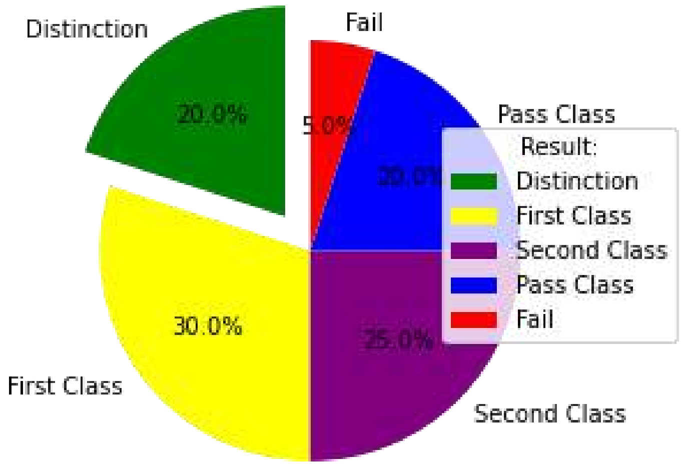
The above code will create pie chart as follow:



We can set the ***explode*** argument in a pie chart to set the fraction of radius with which we offset each wedge.

```
# Importing library
import matplotlib.pyplot as plt

# Labels and area covered by each label
result = ["Distinction", "First Class", "Second Class",
"Pass Class", "Fail"]
area = [20,30,25,20,5]
mycolor = ["Green", "Yellow", "Purple", "Blue", "Red"]
myexplodes = [0.2,0,0,0,0]

# Plotting pie chart
plt.pie(area, labels=result, startangle=90,
colors=mycolor, explode=explodes, autopct='%1.1f%%')
plt.legend(title="Result:", loc="right")
plt.axis('equal')
plt.show()
```

The above code will create pie chart as follow:



2.5 Area Chart

An area chart is similar as a line chart excepts that the area between an x-axis and a line is filled with color or shading. The matplotlib library is used to build the area chart. There are two different functions is used to build an area chart with matplotlib.

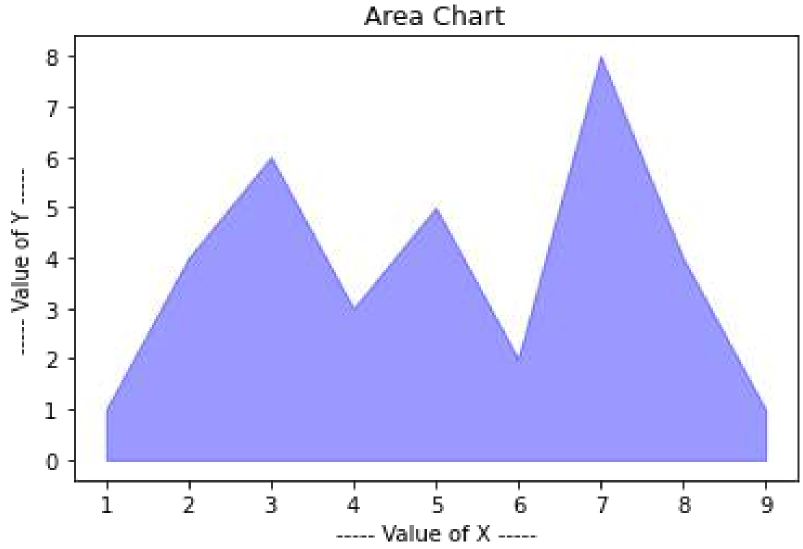
- fill_between() function
- stackplot() function

The ***fill_between()*** function is used to plot an area chart.

Example: Here we take an example of two dataset x and y. The x-axis represents the x values and y-axis represents the y values. Here we also set the chart title, labels on both the axis and ***color*** as a ***Blue***.

```
# Importing library
import matplotlib.pyplot as plt
import numpy as np

# Dataset
x=range(1,10)
y=[1,4,6,3,5,2,8,4,1]

# Titles of chart
plt.title("Area Chart")
plt.xlabel("----- Value of X----- ")
plt.ylabel("----- Value of Y----- ")

# Plotting area chart
plt.fill_between(x,y,color="Blue",alpha=0.4)
plt.show()
```

The above code will create area chart as follow:

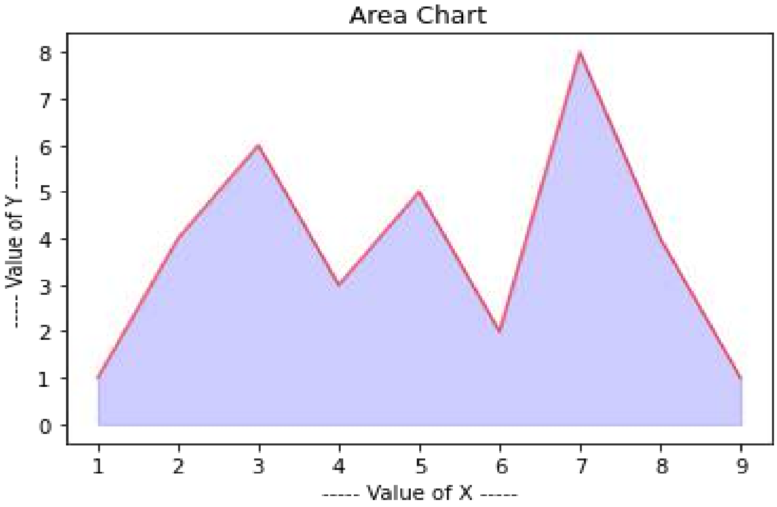


We can change the line color also. We set ***color*** as an argument to change the color. Here we set ***Red*** as a color of line.

```
# Importing library
import matplotlib.pyplot as plt
import numpy as np
```

```
# Dataset
x=range(1,10)
y=[1,4,6,3,5,2,8,4,1]
# Titles of chart
plt.title("Area Chart")
plt.xlabel("----- Value of X----- ")
plt.ylabel("----- Value of Y----- ")
# Plotting area chart
plt.fill_between( x, y, color="Blue", alpha=0.2)
plt.plot(x, y, color="Red", alpha=0.6)
plt.show()
```

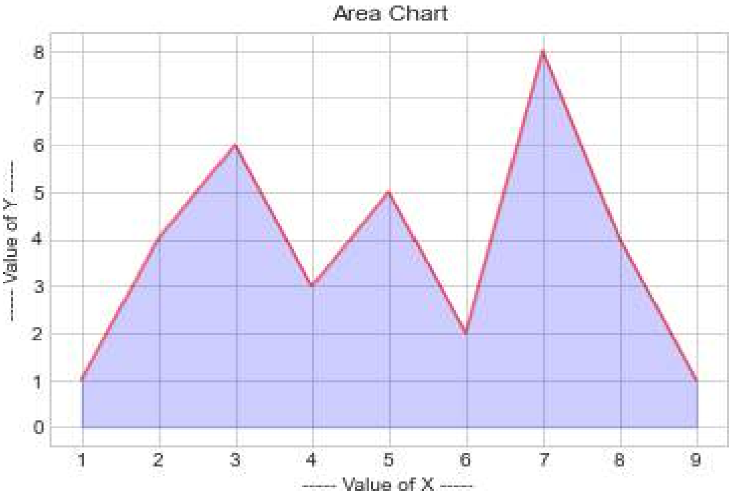
The above code will create area chart as follow:



We can also set and use the plotting style. The various plotting style can be used such as ***fivethirtyeight, seaborn-pastel, seaborn-whitegrid, dark_background*** etc. Here we set ***seaborn-whitegrid*** as a plotting style.

```
# Importing library
import matplotlib.pyplot as plt
import numpy as np
# Dataset
x=range(1,10)
y=[1,4,6,3,5,2,8,4,1]
# Titles of chart
plt.title("Area Chart")
plt.xlabel("----- Value of X----- ")
plt.ylabel("----- Value of Y----- ")
# Plotting style
plt.style.use('dark_background')
# Plotting area chart
plt.fill_between( x, y, color="Blue", alpha=0.2)
plt.plot(x, y, color="Red", alpha=0.6)
plt.show()
```

The above code will create area chart as follow:

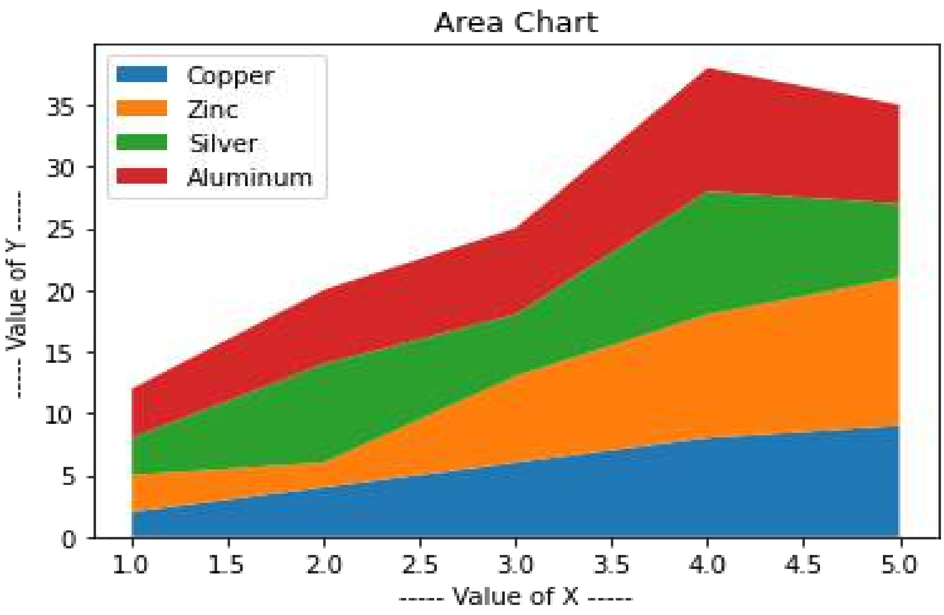


The ***stackplot()*** function is also used to plot an area chart.

Example: Here we take an example of metals such as copper, zinc, silver and aluminum. The x-axis represents the x values and y-axis represents the different metals. Here we also set the chart title, labels on both the axis and chart legend. Here we set ***upper left*** as a location for chart legend.

```
# Importing library
import matplotlib.pyplot as plt
import numpy as np
# Dataset
x=range(1,6)
Copper=[2,4,6,8,9]
Zinc=[3,2,7,10,12]
Silver=[3,8,5,10,6]
Aluminum=[4,6,7,10,8]
# Titles of chart
plt.title("Area Chart")
plt.xlabel("----- Value of X----- ")
plt.ylabel("----- Value of Y----- ")
# Plotting stacked area chart
plt.stackplot(x, Copper, Zinc, Silver, Aluminum,
labels=['Copper','Zinc','Silver','Aluminum'])
plt.legend(loc='upper left')
```

The above code will create area chart as follow:



2.6 Candlestick Chart

The candlestick chart is also known as Japanese chart. This type of charts mostly used for technical analysis in trading for visualizing of price with specified time period. The ***mpl_finance*** module of matplotlib is used to create this chart in Python.

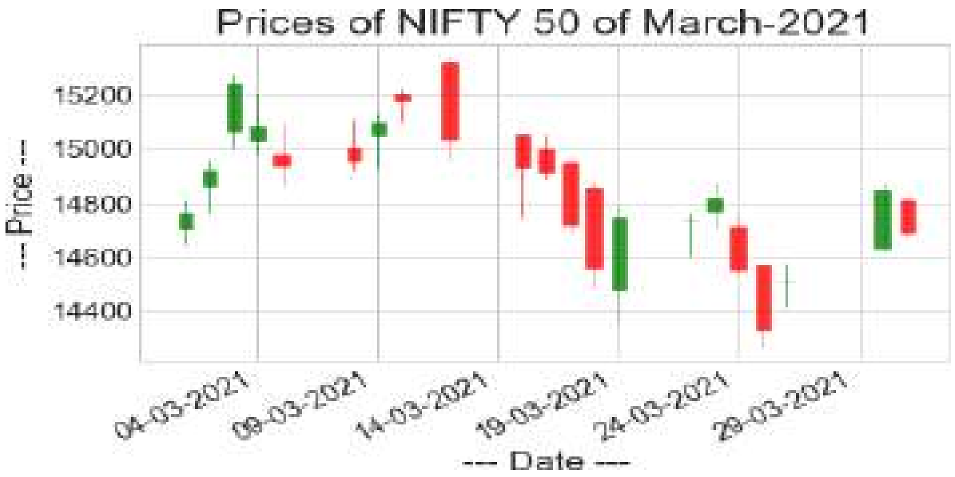
Example: Here we take an example of historical index data of Nifty 50 which was downloaded from NSE website. They have four points such as Open, High, Low and Close (OHLC). Here we set ***colorup*** as a ***green*** and ***colordown*** as a ***red*** argument.

```
# Importing library
import matplotlib.pyplot as plt
from mpl_finance import candlestick_ohlc
import pandas as pd
import matplotlib.dates as mpdates

# Dataset reading and extracting
df = pd.read_csv('Nifty50-March.csv')
df = df[['Date', 'Open', 'High', 'Low', 'Close']]

# Converting datetime object and applying map function
df['Date'] = pd.to_datetime(df['Date'])
df['Date'] = df['Date'].map(mpdates.date2num)

#  Creating  subplots
fig, ax = plt.subplots()

# Plotting candlestick chart
candlestick_ohlc(ax, df.values, width = 0.6, colorup =
'green', colordown = 'red', alpha = 0.8)
ax.grid(True)

# Plotting style
plt.style.use('seaborn-whitegrid')

# Setting title and labels of chart
plt.title('Prices of NIFTY 50 of March-2021')
```

```
ax.set_xlabel('--- Date ---')
ax.set_ylabel('--- Price ---')

# Date formatting
dt_format = mpdates.DateFormatter('%d-%m-%Y')
ax.xaxis.set_major_formatter(dt_format)
fig.autofmt_xdate()

# Show plot
plt.show()
```

Note: The above historical index data of Nifty 50 was downloaded from NSE website.

The above code will create candlestick chart as follow:



2.7 Bubble Chart

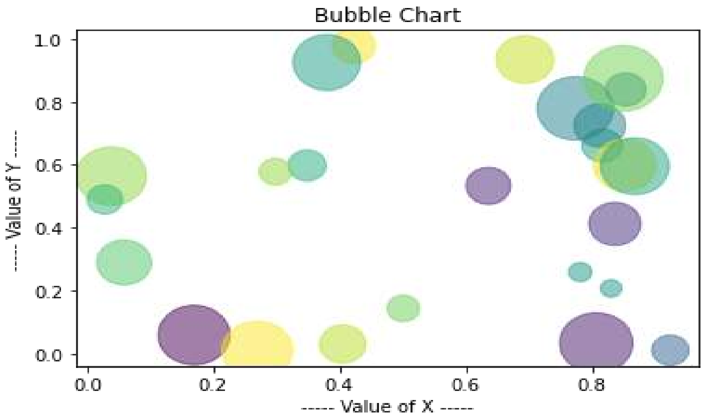
The bubble chart is very similar to the scatter plot. It displays the data as a cluster of circles. The ***scatter()*** function of matplotlib library is used to create a bubble chart in Python. To create bubble chart, it required the data of xy coordinates, size and color of bubbles.

Example: Here we create random numbers for xy coordinates and size of bubbles. Here we set the random colors for bubbles.

```
# Importing library
import matplotlib.pyplot as plt
import numpy as np

# Dataset
x = np.random.rand(25)
y = np.random.rand(25)
z = np.random.rand(25)
colors = np.random.rand(25)

# Titles of chart
plt.title("Bubble Chart")
plt.xlabel("----- Value of X----- ")
plt.ylabel("----- Value of Y----- ")
```

```
# Plotting bubble chart
plt.scatter(x, y, s=z*2000, c=colors, alpha=0.5)
plt.show()
```

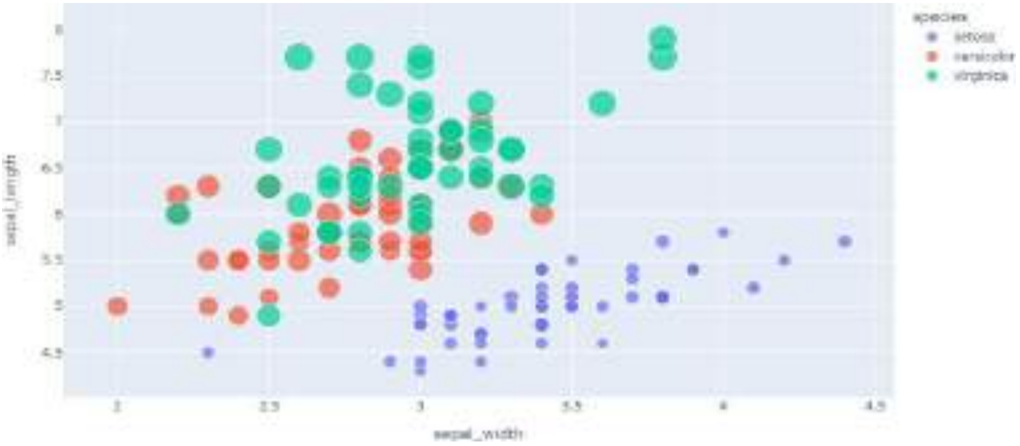
The above code will create bubble chart as follow:



Example: Here we used the ***iris*** dataset to create bubble chart.

```
# Importing libraries
import matplotlib.pyplot as plt
import plotly.express as px

# Create data from iris dataset
df = px.data.iris()

# Plotting bubble chart
plt = px.scatter(df, x="sepal_width", y="sepal_length",
color="species", size='petal_length',
hover_data=['petal_width'])
plt.show()
```

The above code will create bubble chart as follow:



2.8 Surface Plot

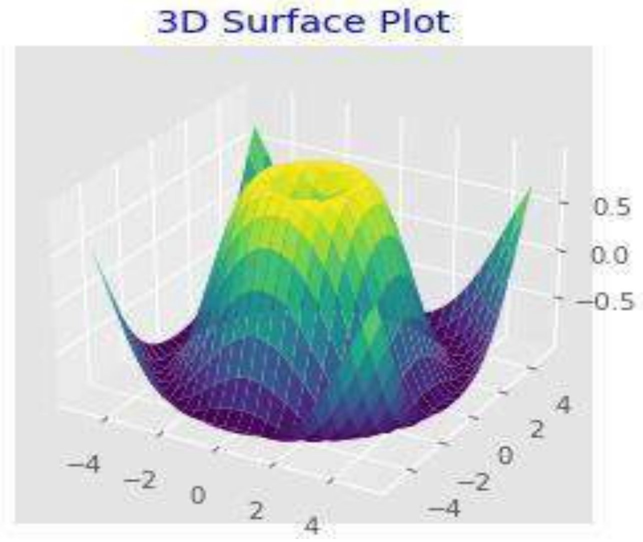
This plot is like as wireframe plot, but each face of the wireframe is filled polygon.

This plot shows the functional relationship between one dependent variable and two independent variables. The ***plot_surface()*** function is used to create a surface plot. Here we create a 3D surface plot.

Example: Here we draw a 3D surface plot.

```
# Importing library
import numpy as np
import matplotlib.pyplot as plt

# 3D projection
fig = plt.figure()
ax = plt.axes(projection="3d")

# Function
def func(x, y):
    return np.sin(np.sqrt(x * x + y * y))

# All three axis
x = np.linspace(-5, 5, 25)
y = np.linspace(-5, 5, 25)
X, Y = np.meshgrid(x, y)
Z = func(X, Y)

# 3D surface plotting
ax.plot_surface(X, Y, Z, rstride=1, cstride=1,
cmap='viridis', edgecolor='none')
ax.set_title("3D Surface Plot", color="blue")
plt.show()
```

The above code will plot as follow:



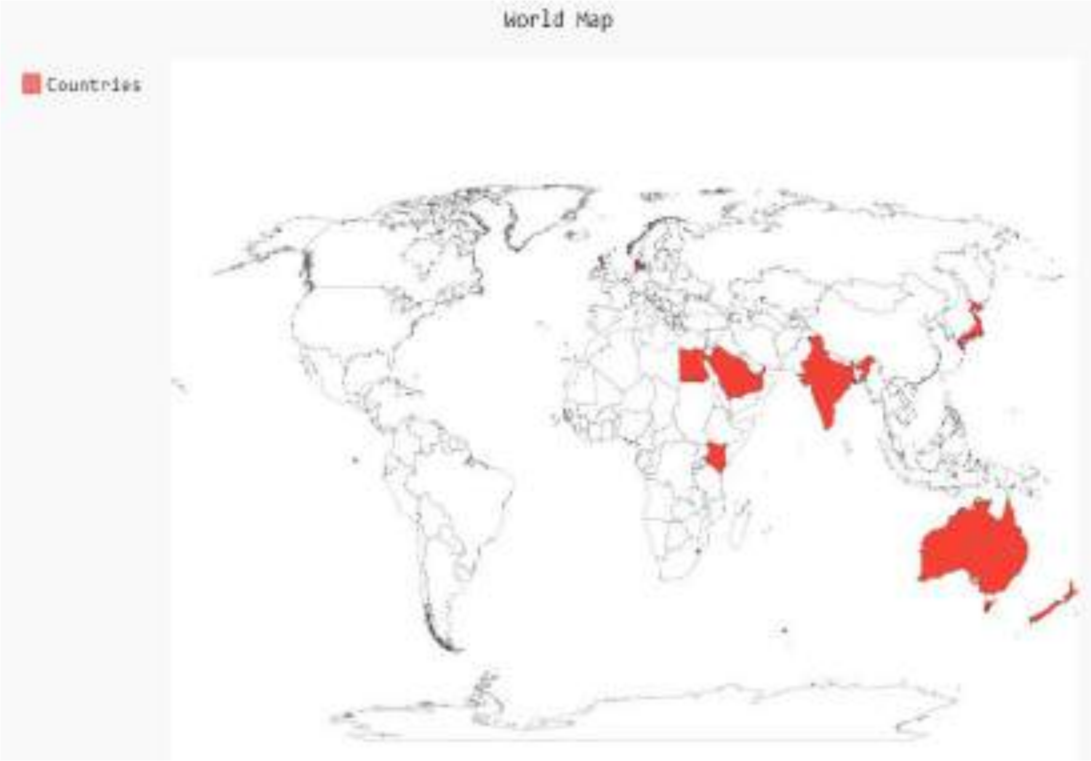
2.9 Map Chart

Map chart is used to represent the different types of information in a context, often geographical using one or more layers. It helps to compare the values and show the categories across geographical regions. The map chart is mostly used to represent geographical regions in a data such as regions, states, countries etc. The map chart contains interactive shapes or display separate markers on an image or map background.

Example: Here we draw a world map of selected country.

```
# Importing library
import pygal

# Create map
wm = pygal.maps.world.World()

# Title of map
wm.title = 'World Map'

# List of countries
wm.add('Countries',
{'ae','au','dk','eg','in','jp','ke','nz','sa','sz'})

# Save the map
wm.render_to_file('map1.svg')
```

The above code will plot world map with highlighted selected countries as follow:

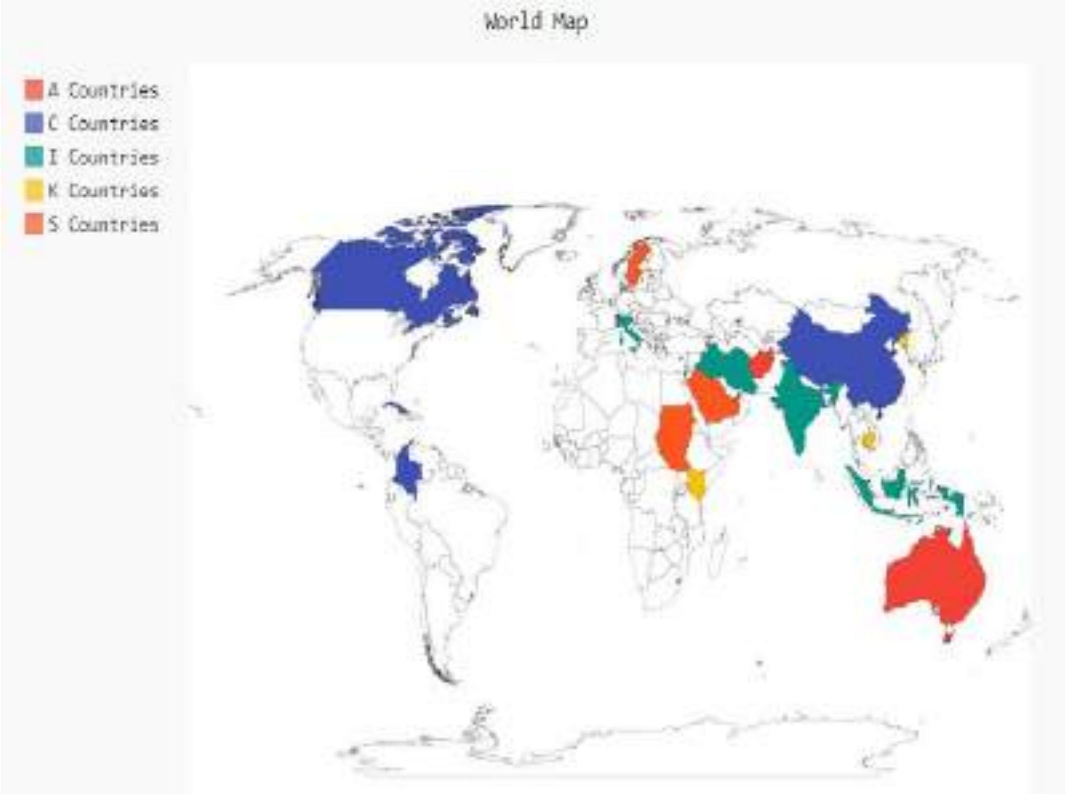


We can also set the different colors for different series of countries. Here we created a world map with selected series of countries as follows:

```
# Importing library
import pygal
```

```
# Create map
wm = pygal.maps.world.World()

# Title of map
wm.title = 'World Map'

# List of countries
wm.add('A Countries', ['ae','af','au'])
wm.add('C Countries', ['ca','ch','cn','co','cu','cy'])
wm.add('I Countries', ['id','il','in','iq','ir','it'])
wm.add('K Countries', ['ke','kh','kp','kw'])
wm.add('S Countries', ['sa','sd','se','sg','sz'])

# Save the map
wm.render_to_file('map2.svg')
```

The above code will plot world map with highlighted selected countries as follow:

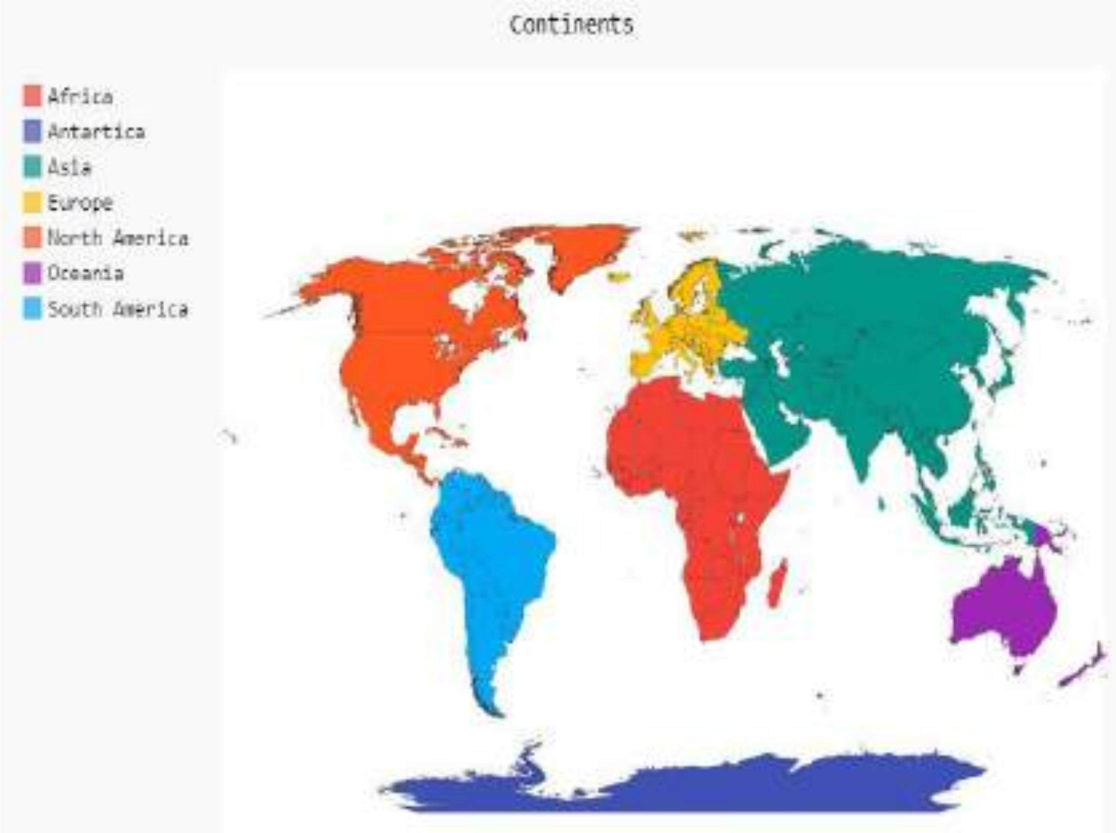


We can also display the different continents in the world. Here we created a world map with continents as follows:

```
# Importing library
import pygal

# Create map
wm = pygal.maps.world.SupranationalWorld()
```

```
# Title of map
wm.title = 'Continents'

# List of continents
wm.add('Africa', [('africa', 1)])
wm.add('Antartica', [('antartica', 1)])
wm.add('Asia', [('asia', 1)])
wm.add('Europe', [('europe', 1)])
wm.add('North America', [('north_america', 1)])
wm.add('Oceania', [('oceania', 1)])
wm.add('South America', [('south_america', 1)])

# Save the map
wm.render_to_file('map3.svg')
```

The above code will plot world map with different continents as follow:



2.10 Infographics

Infographics (information graphics) is a graphical visual representation of data, information or knowledge intended to represent information quickly, clearly and easy to understandable formats. It combines visual imagery, data charts and less text together. It improves the ability of human to see the trends, patterns and relationships in a data.

The reasons to use infographics are:

- Easy to read
- Easy to share
- Easy for marketing
- Easy to summarize content

There are main two basic functions of infographics: static and interactive. The static infographics generally constructed using template. It is very simple solutions when the data and information are remains unchange. The interactive infographics is more complex in terms of designing and programming. It is continuously updating and used to engage audience to give attention and represent the more updated and advances information to the users.

The various types of designing are used for infographics. It is selected based on the types of information that you want to represent clearly. The most common types of infographics are list, comparison, geographic, statistical, informational, data visualization, timeline, interactive, etc.

3. SUMMARY

The students will learn many things in this module such as data mining and data visualization concepts. They will be able to create a various type of charts or plots using Python.

- Ability to do understand the data mining and data visualization concept.
- Ability to plot the various types of charts including data points, line chart, bar chart, pie chart and area chart.
- Ability to plot advance data visualization charts including candlestick chart, bubble chart, surface chart, map chart and infographics.

4. REFERENCES

Books

1. Davy Cielen, Arno D. B. Meysman, Mohamed Ali : Introducing Data Science, Manning Publications Co.
2. Stephen Klosterman (2019) : Data Science Projects with Python, Packt Publishing
3. Jake VanderPlas (2017) : Python Data Science Handbook: Essential Tools for Working with Data, O’Reilly
4. Wes McKinnery and Pandas Development Team (2021) : pandas : powerful Python data analysis toolkit, Release 1.2.3

Web References

1. https://www.oreilly.com
2. https://www.geeksforgeeks.org

3. https://www.tutorialspoint.com
4. https://www.w3schools.com
5. https://pandas.pydata.org
6. https://pbpython.com
7. https://matplotlib.org
8. https://www.logianalytics.com
9. https://seleritysas.com
10. https://www.semrush.com

QUESTIONS

Short Answer:

1. What is chart?
2. List types of charts.
3. What is scatter plot?
4. What is line chart?
5. What is candlestick chart?
6. What is surface plot?
7. What is infographics?

Long Answer:

1. Explain bar chart with example.
2. Explain pie chart with example.
3. Explain area chart with example.
4. Explain bubble chart with example.
5. Explain map chart with example.

PRACTICALS

1. Create and display scatter plot with different arguments.
2. Create and display bar plot with different arguments.
3. Create and display pie plot with different arguments.
4. Create and display area plot with different arguments.
5. Create and display bubble plot with different arguments.
6. Create and display map chart with different arguments.
7. Create and display candlestick plot.
8. Create and display surface plot.

MODULE - VII

VISUALIZING DATA PROGRAMMATICALLY

:: TOPICS ::

- Google chart
- Basics of bar chart
- Basics of pie chart
- Working with chart animation

Dr. Chetan R. Dudhagara

Assistant Professor and Head
Department of Communication & Information Technology
International Agribusiness Management Institute
Anand Agricultural University
Anand, Gujarat, India

OBJECTIVES

The main objective of this module is to understand the data visualization concepts programmatically. The various data visualization techniques and charts such as google charts, bar chart and pie chart are covered in this module.

5. INTRODUCTION

Data science become a buzzword that everyone talks about the data science. Data science is an interdisciplinary field that combines different domain expertise, computer programming skills, mathematics and statistical knowledge to find or extract the meaningful or unknown patterns from unstructured and structure dataset. Data science is useful for extraction, preparation, analysis and visualization of various information. Various scientific methods can be applied to get insight in data.

Data visualization is a graphical representation of data and quantitative information by using various graphical tools such as graphs and charts. It is more useful to understand the trends, patterns and outliers in the data.

6. DATA VISULIZATION

Charts is the representation of data in a graphical form. It helps to summarizing and presenting a large amount of data in a simple and easy to understandable formats. By placing the data in a visual context, we can easily detect or identify the patterns, trends and correlations among them.

Python provides various easy to use multiple graphics libraries for data visualization. These libraries are work with both small and large datasets.

Python has multiple graphics libraries with different features. Some of the most popular and commonly used Python data visualization libraries are Matplotlib, Pandas, Seaborn, Plotly and ggplot.

Google is also providing very powerful tool for visualization which is called google chart. It is very simple, user-friendly and free tool for creating interactive and animated charts.

Google chart API (Application Programming Interface) is an extremely easy and simple tool to create a chart from the data and embed it into a webpage. It has numerous functions and properties for creating efficient and interactive charts. It has good cross-browser compatibility and also supports legacy browsers well.

7. STARTING WITH GOOGLE CHARTS

Google chart is a very popular library available for charting tool. Initially it was developed and used for its internal applications for rendering charts. Later it was released for the public use also. It helps and used to create variety of charts such as bar chart, pie chart, bubble chart, geo chart, etc.

Google charts is a combination of APIs:

- Google Chart API
- Google Visualization API

- The google chart API is used to create static visualizations from the data and embeds it into a webpage. Basic HTML programming knowledge is recommended to create chart using google chart API. The various types of charts including scatter, line, bar, pie, etc. can be created using it and embedded into webpages.
- The google visualization API is used to creates dynamic visualizations which allow to user interaction within the webpage. The charts are created using java scripting language. The various types of charts including timelines, heat maps, tree maps, etc. can be created using it and embedded into webpages.

3.1 Google Charts API

Google charts is a pure JavaScript based charting library to enhance web application by adding interactive charting capability. It provides variety of chart such as bar chart, pie chart, line chart, spline chart, area chart, etc.

There are two ways to use google charts

- Using downloaded google chart
- Using content delivery network (CDN)

Using downloaded google chart:

- Include the google charts JavaScript file in the HTML page using following script.
- **Syntax:**

```
<head>
                    <script src = “googlecharts/loader.js/” >
</head>
```

Using content delivery network:

- Include the google charts JavaScript file in the HTML page using following script.
- **Syntax:**

```
<head>
                    <script src = “https://www.gstatic.com/charts/loader.js/” >
</head>
```

3.2 Bar Chart

A bar chart or bar graph that presents categorical data using rectangular bars or columns with different heights or lengths proportional to the value that they represent. The bars can be plotted horizontally or vertically. The vertical bar chart is called as column chart also.

To create google bar chart in web page, the following link is added.

```
<script
src = "https://www.gstatic.com/charts/loader.js">
</script>
```

The <div> tag or element is used to display the chart in a web page.

```
<div id="myChart" style="width:700px; height:400px">
</div>
```

The <div> tag or element must have a unique name or id.

To load the google graph API, the ***corechart*** package is use. The callback function is use to call when the API is loaded.

```
google.charts.load('current',{packages:['corechart']});
google.charts.setOnLoadCallback(drawChart);
```

Example: To create a bar chart using google chart of the following data.

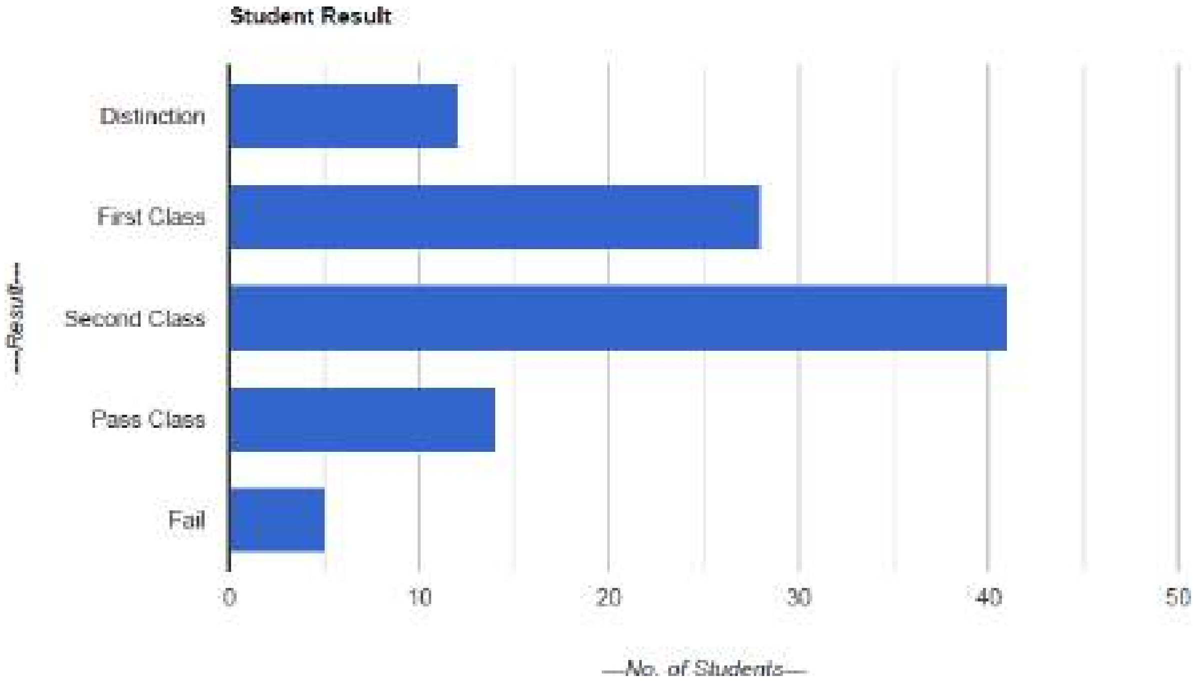
Class	Distinction	First Class	Second Class	Pass Class	Fail
No. of Student	12	28	41	14	5

The above data contains results of 100 students.

The following code will create a bar chart using JavaScript.

```
<html>
  <head>
    <script type="text/javascript"
     src="https://www.gstatic.com/charts/loader.js"></script>
    <script type="text/javascript">
     google.charts.load('current', {packages:['corechart']});
     google.charts.setOnLoadCallback(drawChart);
     function drawChart()
     {
        var data = google.visualization.arrayToDataTable
        ([
          ['Result', 'No. of Students'],
          ['Distinction', 12],
          ['First Class', 28],
          ['Second Class', 41],
          ['Pass Class', 14],
          ['Fail', 5]
        ]);
        var options =
        {
            title: 'Student Result',
            hAxis: {title: '---No. of Students---'},
            vAxis: {title: '---Result---'}
        };
        var chart = new
        google.visualization.BarChart(document.getElementById
        ('barchart'));
        chart.draw(data, options);
      }
    </script>
  </head>
  <body>
    <div id="barchart" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

The above code will create bar chart as follow:



Example: To create a grouped bar chart using google chart of the following data.

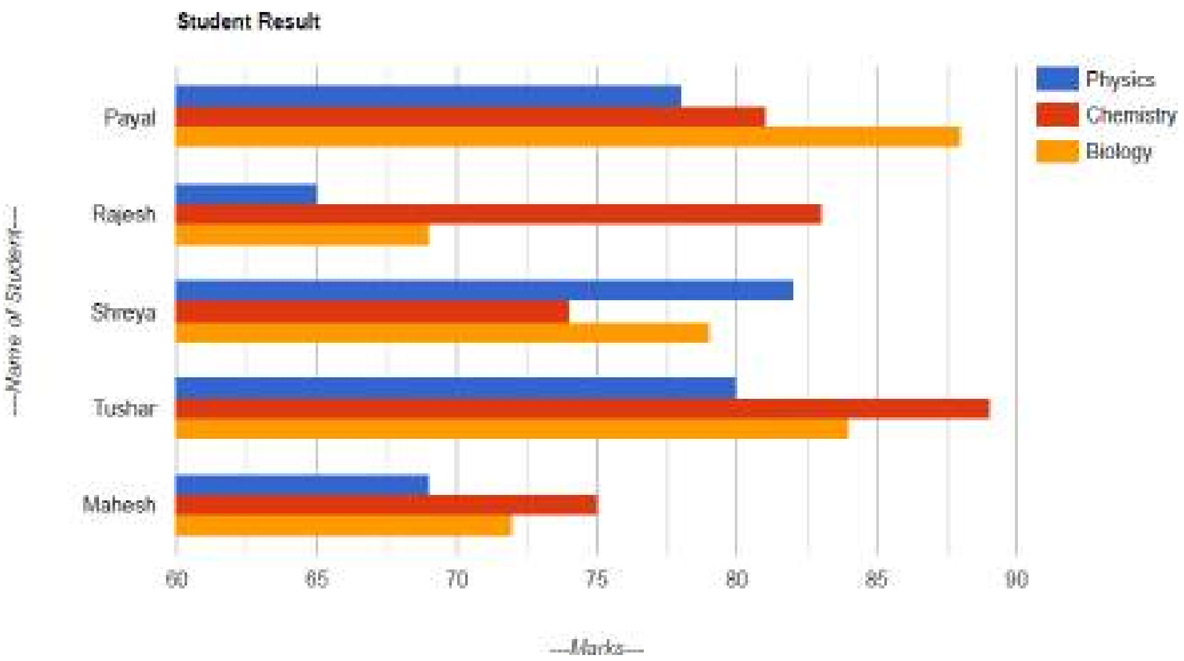
Name of Student	Payal	Rajesh	Shreya	Tushar	Mahesh
Physics	78	65	82	80	69
Chemistry	81	83	74	89	75
Biology	88	69	79	84	72

The above data contains results of three subjects such as physics, chemistry and biology score of 5 students.

The following code will create a grouped bar chart using JavaScript.

```
<html>
  <head>
    <script type="text/javascript"
     src="https://www.gstatic.com/charts/loader.js"></script>
    <script type="text/javascript">
     google.charts.load('current', {packages:['corechart']});
     google.charts.setOnLoadCallback(drawChart);
     function drawChart()
     {
        var data = google.visualization.arrayToDataTable
        ([
          ['Name', 'Physics', 'Chemistry', 'Biology'],
          ['Payal', 78, 81, 88],
          ['Rajesh', 65, 83, 69],
          ['Shreya', 82, 74, 79],
          ['Tushar', 80, 89, 84],
          ['Mahesh', 69, 75, 72]
        ]);
        var options =
        {
            title: 'Student Result'
            hAxis: {title: '---Marks---'},
            vAxis: {title: '---Name of Student---'}
        };
        var chart = new
        google.visualization.BarChart(document.getElementById
        ('barchart'));
        chart.draw(data, options);
      }
    </script>
  </head>
  <body>
    <div id="barchart" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

The above code will create grouped bar chart as follow:



3.3 Pie Chart

A pie chart is a type of graph which display the data in a circular graph which is also divided into slices. The slices of pie chart represent the relative size or quantity of data. It requires a list of categorical values and corresponding numerical values. The term pie represents a whole and a slice represent the portions or parts of a whole.

Example: To create a pie chart using google chart of the following data.

Name of Browser	Market Share (in %)
Chrome	64.67
Safari	19.06
Edge	3.99
Firefox	3.66
Samsung Internet	2.81
Opera	2.36
UC Browser	0.97
Android	0.57
IE	0.5
Other	1.41

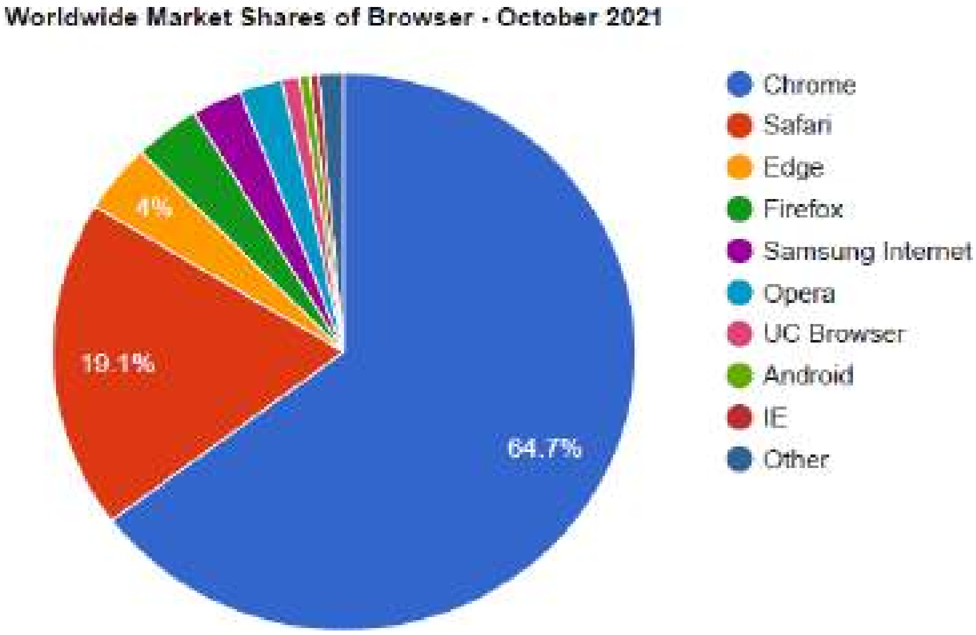
The above data contains worldwide market shares of browsers in October 2021

(Source: https://gs.statcounter.com/browser-market-share)

The following code will create a pie chart using JavaScript.

```
<html>
  <head>
    <script type="text/javascript"
     src="https://www.gstatic.com/charts/loader.js"></script>
    <script type="text/javascript">
     google.charts.load('current', {packages:['corechart']});
     google.charts.setOnLoadCallback(drawChart);
     function drawChart()
     {
        var data = google.visualization.arrayToDataTable
        ([
          ['Browser', 'Market Share'],
          ['Chrome', 64.67],
          ['Safari', 19.06],
          ['Edge', 3.99],
          ['Firefox', 3.66],
          ['Samsung Internet', 2.81],
          ['Opera', 2.36],
          ['UC Browser', 0.97],
          ['Android', 0.57],
          ['IE', 0.5],
          ['Other', 1.41]
        ]);
        var options =
        {
          title: 'Worldwide Market Shares of Browser - October 2021'
        };
        var chart = new google.visualization.PieChart
          (document.getElementById('piechart'));
        chart.draw(data, options);
      }
    </script>
  </head>
  <body>
    <div id="piechart" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

The above code will create pie chart as follow:



Example: To create a donut chart using google chart of the following data.

Activity	Sleep	School	Study	Eat	Sports	Entertainment	Refresh
No. of Hour	8	6	2	2	2.5	2	1.5

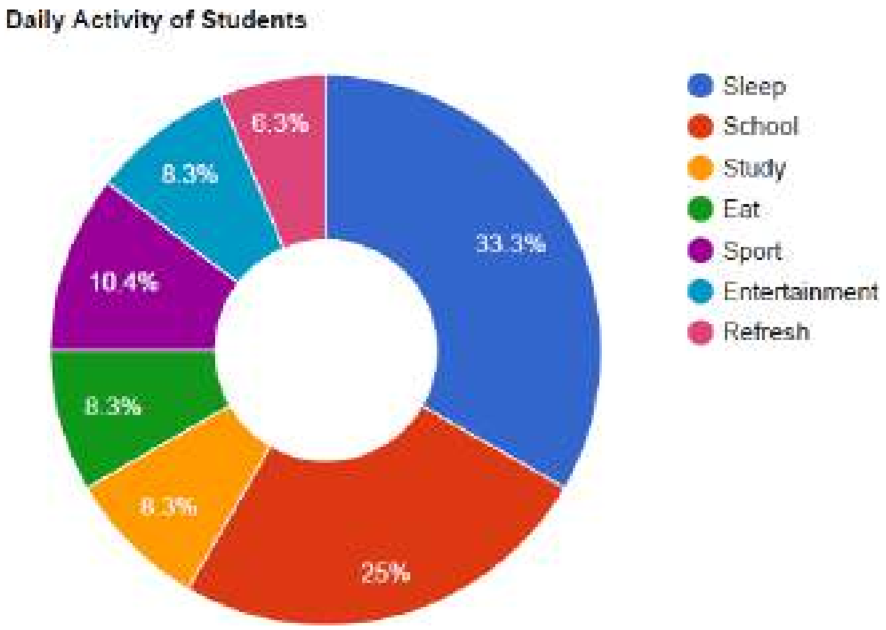
The above data contains numbers of hour spent in daily activities by the students.

The following code will create a donut chart using JavaScript.

```
<html>
  <head>
    <script type="text/javascript"
     src="https://www.gstatic.com/charts/loader.js"></script>
    <script type="text/javascript">
     google.charts.load('current', {packages:['corechart']});
     google.charts.setOnLoadCallback(drawChart);
     function drawChart()
     {
        var data = google.visualization.arrayToDataTable
        ([
          ['Activity', 'No. of Hours'],
          ['Sleep', 8],
          ['School', 6],
          ['Study', 2],
          ['Eat', 2],
          ['Sport', 2.5],
          ['Entertainment', 2],
          ['Refresh', 1.5],
        ]);
        var options =
        {
          title: 'Daily Activity of Students', pieHole: 0.4
        };

        var chart = new google.visualization.PieChart
```

```
         (document.getElementById('piechart'));
        chart.draw(data, options);
      }
    </script>
  </head>
  <body>
    <div id="piechart" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

The above code will create donut chart as follow:



Example: To create a 3D pie chart using google chart of the following data.

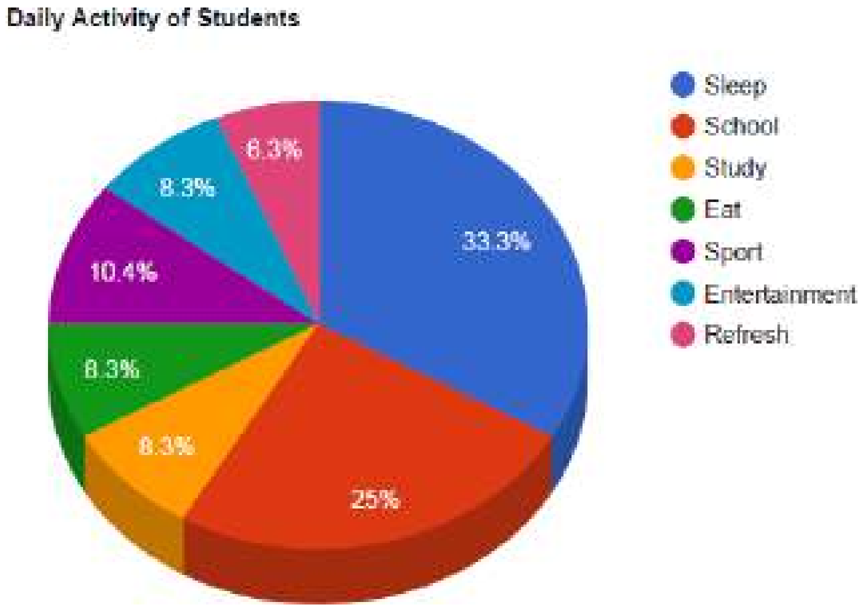
Activity	Sleep	School	Study	Eat	Sports	Entertainment	Refresh
No. of Hour	8	6	2	2	2.5	2	1.5

The above data contains numbers of hour spent in daily activities by the students.

The following code will create a 3D pie chart using JavaScript.

```
<html>
  <head>
    <script type="text/javascript"
     src="https://www.gstatic.com/charts/loader.js"></script>
    <script type="text/javascript">
     google.charts.load('current', {packages:['corechart']});
     google.charts.setOnLoadCallback(drawChart);
     function drawChart()
     {
        var data = google.visualization.arrayToDataTable
        ([
          ['Activity', 'No. of Hours'],
          ['Sleep', 8],
          ['School', 6],
          ['Study', 2],
          ['Eat', 2],
          ['Sport', 2.5],
          ['Entertainment', 2],
          ['Refresh', 1.5],
        ]);
        var options =
        {
          title: 'Daily Activity of Students', is3D:true
        };
        var chart = new google.visualization.PieChart
          (document.getElementById('piechart'));
        chart.draw(data, options);
      }
    </script>
  </head>
  <body>
    <div id="piechart" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

The above code will create 3D pie chart as follow:



8. WORKING WITH CHART ANIMATIONS

The charts which represents data in a dynamic and animated way for better understanding is called animation chart. The google charts can animated smoothly in two ways, either on startup when you draw the chart at first time or when you redraw the chart after making changes on data or operations. The animated chart displays several chart states one after the other. The animation chart required a little additional efforts and knowledge to create dynamic changes and movements in a chart.

Example: To create an animated bar chart using google chart of the following data.

Product	Computer	Printer	Tablet	Mobile	Web Camera	Head Phone
No. of Item Sales	12	8	22	40	25	48

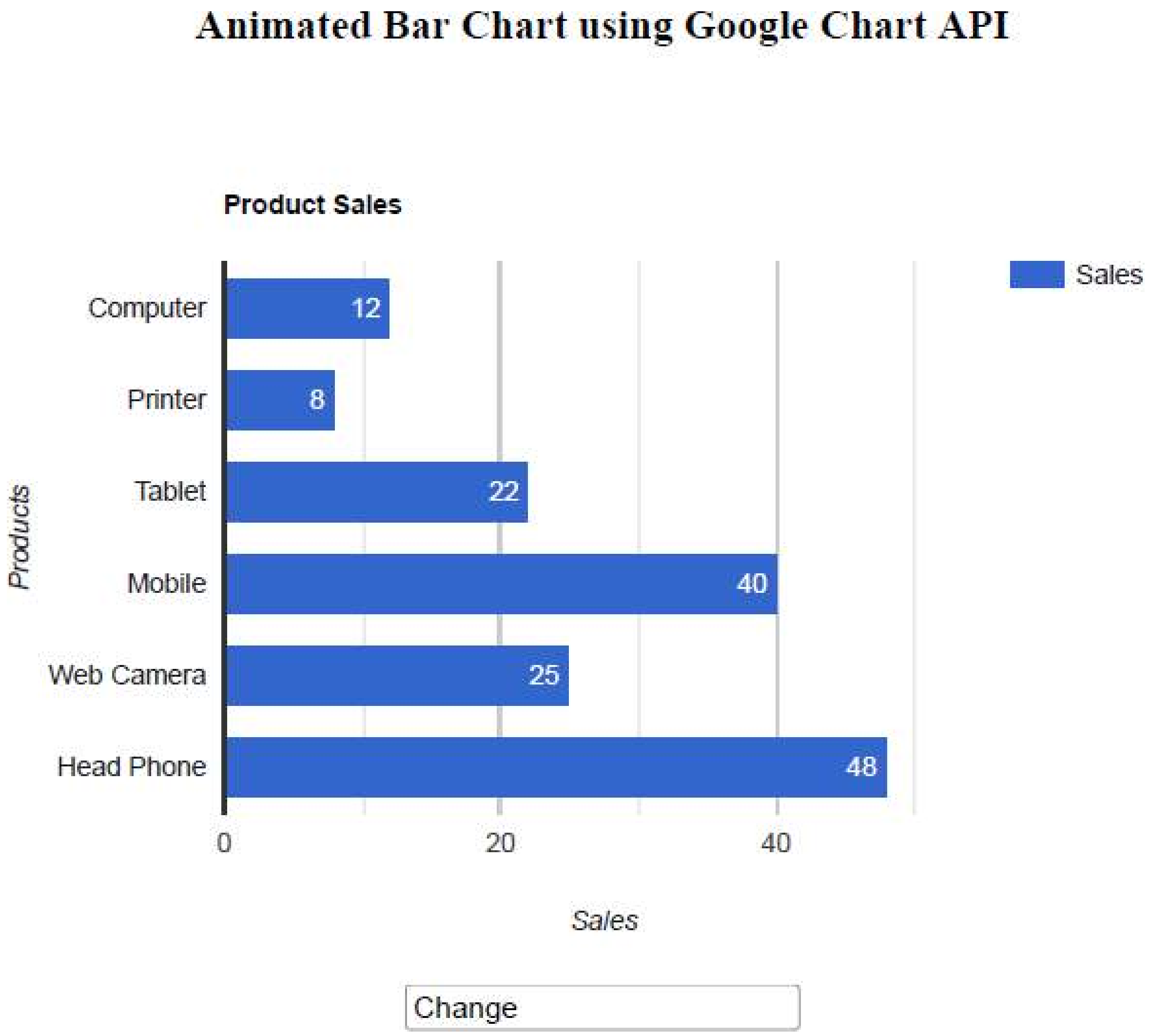
The above data contains sales of different computer related items.

The following code will create an animated bar chart using JavaScript.

```
<!DOCTYPE html>
<html>
<head>
<center><h3> Animated Bar Chart using Google Chart API </h3>
<script type="text/javascript" src="https://www.google.com/jsapi">
</script>
   <script type="text/javascript">
   google.load('visualization', '1', {packages:['corechart']});
   </script>
   <script type="text/javascript">
   function renderChart()
   {
      var data = google.visualization.arrayToDataTable
      ([
           ['Product', 'Sales', { role: 'annotation' } ],
           ['Computer', 0, "12"],
           ['Printer', 0, "8"],
           ['Tablet', 0, "22"],
           ['Mobile', 0, "40"],
           ['Web Camera', 0, "25"],
           ['Head Phone', 0, "48"]
      ]);
      var options =
      {
           title : "Product Sales",
           hAxis: { title: "Sales", viewWindow: { min: 0, max: 55 } },
           vAxis: { title: "Products" },
           animation: {duration: 1000, easing: 'inAndOut',}
      };
      var button = document.getElementById('changeData');
      var initialAnimationPlayed = false;
      var chart = new google.visualization.BarChart(
           document.getElementById("chart"));
      google.visualization.events.addListener(chart, 'ready',
      function()
      {
        if (!initialAnimationPlayed)
        {
           initialAnimationPlayed = true;
           data.setValue(0, 1, 12);
           data.setValue(1, 1, 8);
           data.setValue(2, 1, 22);
           data.setValue(3, 1, 40);
           data.setValue(4, 1, 25);
           data.setValue(5, 1, 48);
           chart.draw(data, options);
         }
         });
         chart.draw(data, options);
         var firstData = true;
         button.onclick = function ()
         {
           if (!firstData)
           {
             firstData = !firstData;
             data.setValue(0, 1, 12);
             data.setValue(1, 1, 8);
             data.setValue(2, 1, 22);
             data.setValue(3, 1, 40);
             data.setValue(4, 1, 25);
             data.setValue(5, 1, 48);
             data.setValue(0, 2, "12");
             data.setValue(1, 2, "8");
             data.setValue(2, 2, "22");
             data.setValue(3, 2, "40");
             data.setValue(4, 2, "25");
             data.setValue(5, 2, "48");
           }
           else
           {
             firstData = !firstData;
             data.setValue(0, 1, 40);
             data.setValue(1, 1, 25);
             data.setValue(2, 1, 15);
             data.setValue(3, 1, 30);
             data.setValue(4, 1, 10);
             data.setValue(5, 1, 18);
             data.setValue(0, 2, "40");
             data.setValue(1, 2, "25");
             data.setValue(2, 2, "15");
             data.setValue(3, 2, "30");
             data.setValue(4, 2, "10");
             data.setValue(5, 2, "18");
           }
           chart.draw(data, options);
         };
     }
     google.setOnLoadCallback(renderChart);
</script>
<div id="chart" style="width:550px; height:400px; margin: 0 auto">
</div>
</head>
<body>
<div>
       <input id="changeData" style="Button" value="Change">
</div>
</body>
```

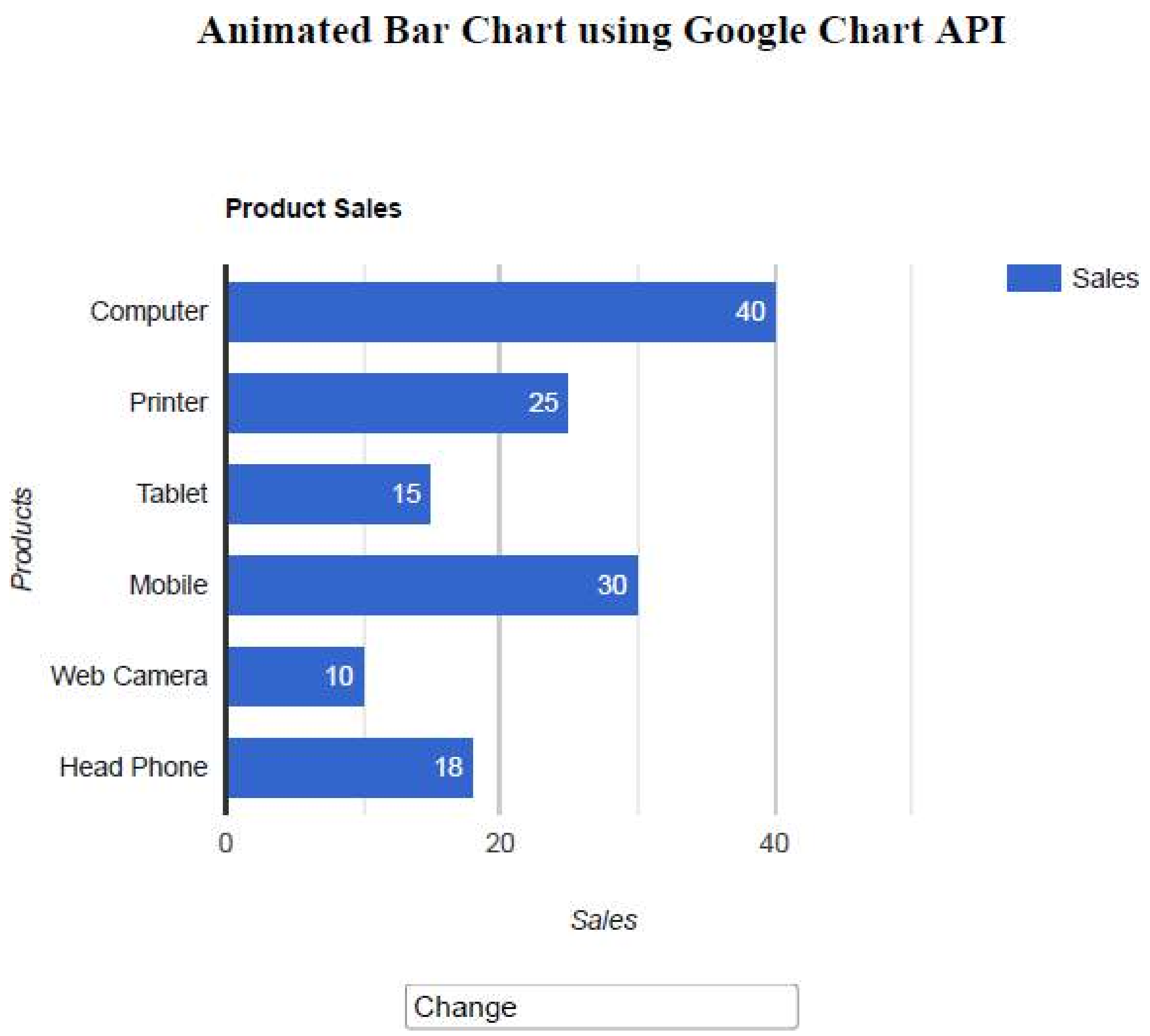
```
</html>
```

The above code will create animated bar chart as follow:



Animated Bar Chart using Google Chart API

After the click on change button the bar chart is animated as follows:



Animated Bar Chart using Google Chart API

9. SUMMARY

The students will learn data visualization concepts using google chart API in JavaScript. They will be able to create a various type of charts or plots using google chart API.

- Ability to do understand the data visualization concept.
- Ability to plot the various types of charts including bar chart and pie chart with animation using google chart.

10. REFERENCES

Books

1. Jon J. Raasch, Graham Murray, Vadim Ogievetsky, Joseph Lowery (2015) : JavaScript® and jQuery® for Data Analysis and Visualization, Published by John Wiley & Sons, Inc.
2. Stephen Klosterman (2019) : Data Science Projects with Python, Packt Publishing.
3. Jake VanderPlas (2017) : Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly.

Web References

11. https://developers.google.com/chart/
12. https://www.encodedna.com/google-chart/create-interactive-graphs-using-google-chart.htm
13. https://www.tutorialspoint.com/googlecharts/googlecharts_pie_basic.htm
14. https://old.dataone.org/software-tools/google-charts
15. https://codeactually.com/googlecharts.html
16. https://www.w3schools.com/whatis/whatis_google_charts.asp
17. https://en.wikipedia.org/wiki/Usage_share_of_web_browsers
18. https://gs.statcounter.com/browser-market-share
19. https://canvasjs.com/javascript-charts/animated-chart/

QUESTIONS

Short Answer:

8. What is chart?
9. List types of charts.
10. What is google chart API?
11. What is content delivery network?
12. What is bar char?
13. What is pie chart?

Long Answer:

6. Explain google chart in detail.
7. Explain bar chart using google chart with example.
8. Explain pie chart using google chart with example.
9. Explain animated chart using google chart with example.

PRACTICALS

9. Create bar chart using google chart.
10. Create pie chart using google chart.
11. Create animated chart using google chart.