**R.G. KEDIA COLLEGE OF COMMERCE**
**DEPARTMENT OF BUSINESS MANAGEMENT**

**SUBJECT SYNOPSIS**

**Subject: Business Analytics** **Paper Code: MB-304-S-II**
**Year /Semester: II/III**

**Academic Year: 2023-24**
**SYLLABUS**

**Course Objectives:**
1. The objective is to provide knowledge of data science
2. To provide basic statistical tools
3. State the importance of data in current business scenario
4. To develop contingent business models for better analysis

**Course Outcomes**:
1. Students can use data as tool for business analysis
2. The basic statistics provides a road map to learners
3. Micro metrics makes the students to identify data gaps

## Unit – I: Introduction to Business Analytics:

Definition of Business Analytics, Categories of Business Analytical methods and models, Business Analytics in practice, Big Data - Overview of using Data, Types of Data- Business decision modeling.

## Unit – II: Descriptive Analytics:

Overview of Description Statistics (Central Tendency, Variability), Data Visualization - Definition, Visualization Techniques – Tables, Cross Tabulations, charts, Data Dashboards using Advanced Ms-Excel or SPSS.

## Unit – III: Predictive Analytics:

Trend Lines, Regression Analysis – Linear & Multiple, Predictive modeling, forecasting Techniques, Data Mining - Definition, Approaches in Data Mining- Data Exploration & Reduction, Data mining and business intelligence, Data mining for business Classification, Association, Cause Effect Modeling.

## Unit – IV: Prescriptive Analytics:

Overview of Linear Optimization, Non Linear Programming Integer Optimization, Cutting Plane algorithm and other methods, Decision Analysis – Risk and uncertainty methods - Text analytics Web analytics.

## Unit – V: Programming Using R:

R Environment, R packages, Reading and Writing data in R, R functions, Control Statements, Frames and Subsets, Managing and Manipulating data in R.

## Suggested Books:


1. Camm, Cochran, Fry, Ohlmann, Anderson, Sweeney, Williams - **Essentials of Business Analytics,** Cengage Learning.
2. James Evans, **Business Analytics,** Pearson, Second Edition, 2017.
3. Albright Winston, **Business Analytics - Data Analysis - Data Analysis and Decision Making,** Cengage Learning, Reprint 2016.
4. Sahil Raj**, Business Analytics**, Cengage Learning.

**R.G. KEDIA COLLEGE OF COMMERCE**
**DEPARTMENT OF BUSINESS MANAGEMENT**
**SUBJECT SYNOPSIS**

**Subject: Business Analytics** **Paper Code: MB-304-S-II**
**Year /Semester: II/III**

## Unit – I: Introduction to Business Analytics:

Definition of Business Analytics, Categories of Business Analytical methods and models, Business Analytics in practice, Big Data - Overview of using Data, Types of Data- Business decision modeling.

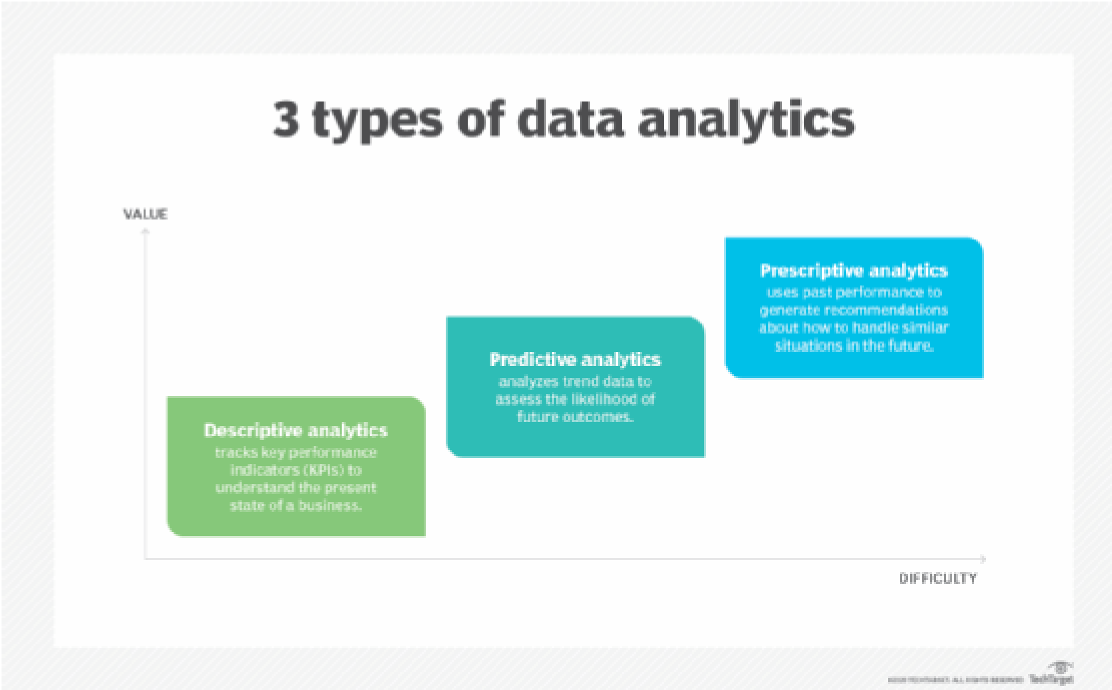
### Definition of Business Analytics:

Business analytics (BA) is a set of disciplines and technologies for solving business problems using data analysis, statistical models and other quantitative methods. It involves an iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis, to drive decision-making.
Data-driven companies treat their data as a business asset and actively look for ways to turn it into a competitive advantage. Success with business analytics depends on data quality, skilled analysts who understand the technologies and the business, and a commitment to using data to gain insights that inform business decisions.

### Types of Business Analytics

Different types of business analytics include the following:

- Descriptive analytics, which tracks key performance indicators (KPIs) to understand the present state of a business;
- Predictive analytics, which analyzes trend data to assess the likelihood of future outcomes; and
- Prescriptive analytics, which uses past performance to generate recommendations for handling similar situations in the future.




### Common challenges of business analytics

Businesses might encounter both business analytics and business intelligence challenges when trying to implement a business analytics strategy:

- **Too many data sources.** There is an increasingly large spectrum of internet-connected devices generating business data. In many cases, they are generating different types of data that must be integrated into an analytics strategy. However, the more complex a data set becomes, the harder it is to use it as part of an analytics framework.
- **Lack of skills.** The demand for employees with the data analytic skills necessary to process BA data has grown. Some businesses, particularly small and medium-sized businesses (SMBs), may have a hard time hiring people with the BA expertise and skills they need.
- **Data storage limitations.** Before a business can begin to decide how it will process data, it must decide where to store it.

### Roles and responsibilities in business analytics

Business analytics professionals' main responsibility is to collect and analyze data to influence strategic decisions that a business makes. Some initiatives they might provide analysis for include the following:

- identifying strategic opportunities from data patterns;
- identifying potential problems facing the business and solutions;
- creating a budget and business forecast;
- monitoring progress with business initiatives;
- reporting progress on business objectives back to stakeholders;
- understanding KPIs; and
- understanding regulatory and reporting requirements.

**The Benefits of Business Analytics**

To club in one phrase: Business Analytics brings actionable insights for businesses. However, here are the main benefits of Business Analytics:

1. Improve operational efficiency through their daily activities.
2. Assist businesses to understand their customers more precisely.
3. Business uses data visualization to offer projections for future outcomes.
4. These insights help in decision making and planning for the future.
5. Business analytics measures performance and drives growth.
6. Discover hidden trends, generate leads, and scale business in the right direction.

**Difference Between Business Intelligence and Business Analytics**

Business Intelligence(BI) uses the past and present to identify trends and patterns in the organizational procedures, while Business Analytics determines the reasons and factors that led to present situations. Business Intelligence focuses mainly on descriptive analysis, while Business Analytics deals with predictive analysis. BI tools are part of Business Analytics that helps understand the Business Analytics process better.

**Business Analytics Examples and Tools**

Many business analytics and business intelligence tools can automate advanced data analytics tasks. Here are a few examples of commercial business analytics software:

- Knime Analytics Platform includes machine learning and high-performance data pipelining

- Dundas Business Intelligence has automated trend forecasting and an intuitive interface
- Qlik's QlikView has data visualization and automated data association features
- Sisense is renowned for its data warehousing and dynamic text-analysis capabilities
- Splunk comes with a user-friendly interface and data visualization capabilities
- Tableau offers sophisticated capabilities for natural language processing and unstructured text analysis.
- Tibco Spotfire is an automated statistical and unstructured text analysis tool with powerful abilities.

Organizations should consider the following factors when choosing a business analytics tool:

- The sources from which their data gets derived
- The kind of data that requires the analysis
- The tool's usability

**A Career in Business Analytics**

The role of Business Analytics professionals may change accordingly to meet organizational goals and objectives. Several individual profiles are closely associated with business analytics when dealing with data.

In this competitive age, business analytics has revolutionized the procedures to discover intelligent insights and gain more profits using their existing methods only. Business Analytics Techniques also help organizations personalize customers with more optimized services and even include their feedback to create more profitable products. Large organizations today are now competing to stay top in the markets by utilizing practical business analytics tools.

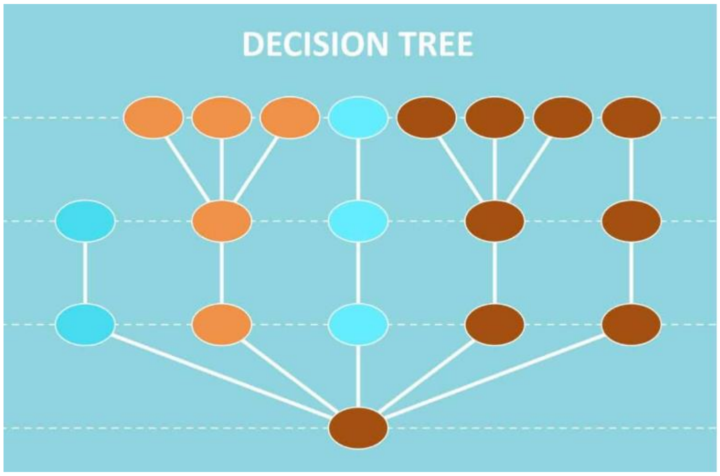
Several business analytics tools are available in the market that offers specific solutions to match requirements. Professionals might need business analytics skills, like understanding and expertise of statistics or SQL to manage them.

**What is Big Data?**

Big data refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods. The act of accessing and storing large amounts of information for analytics has been around for a long time. But the concept of big data gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three V's

| | |
|---|---|
| **Volume** | The amount of data matters. With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes. |
| **Velocity** | Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action. |
| **Variety** | Variety refers to the many types of data that are available. Traditional data types were structured and fit neatly in a relational database. With the rise of big data, data comes in new unstructured data types. Unstructured and semistructured data types, such as text, audio, and video, require additional preprocessing to derive meaning and support metadata. |

**Introduction to Decision Modeling**




The ability to make well-informed decisions is a cornerstone of success, especially in the dynamic and complex world of modern business operations.

Making decisions is a huge part of conducting any business, no matter how big or small it is. Some of these decisions are strategic and have a massive impact on the future of the entire organisation.

On the other hand, some decisions are more routine, repeatable, and are made almost every day during regular business operations. Nevertheless, these everyday decisions still carry great weight and importance and, due to their frequency, are essential in keeping the business processes efficient and effective.

However, with the vast amount of data involved, many different possibilities, and increased demands for process optimisation, making these sorts of decisions and making sure that they are the right ones can be exceptionally challenging.

One of the main responsibilities of business analysis is to help organisations improve their decision-making process, decipher complexities, and drive effective choices that will help companies maintain their operations at an optimum level.

Probably the best way to go about this is by using decision modelling, a potent technique that helps analysts unravel the intricacies of decision-making within an organisation and identify the key steps of this process, as well as develop a systematic approach to understanding and organising data so that important operational decisions can be made easily, quickly, and accurately.

**Subject: Business Analytics** **Paper Code: MB-304-S-II**
**Year /Semester: II/III**

## Unit – II: Descriptive Analytics:

Overview of Description Statistics (Central Tendency, Variability), Data Visualization - Definition, Visualization Techniques – Tables, Cross Tabulations, charts, Data Dashboards using Advanced Ms-Excel or SPSS.

### What Are Descriptive Statistics?

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

KEY TAKEAWAYS

- Descriptive statistics summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the center of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).
- Measures of frequency distribution describe the occurrence of data within the data set (count).

### Types of Descriptive Statistics

All descriptive statistics are either measures of central tendency or measures of variability, also known as measures of dispersion.

### Central Tendency

Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables and general discussions to help people understand the meaning of the analyzed data.

Measures of central tendency describe the center position of a distribution for a data set. A person analyzes the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analyzed data set.

**Measures of Variability**

Measures of variability (or the measures of spread) aid in analyzing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.

So while the average of the data maybe 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

**Distribution**

Distribution (or frequency distribution) refers to the quantity of times a data point occurs. Alternatively, it is the measurement of a data point failing to occur. Consider a data set: male, male, female, female, female, other. The distribution of this data can be classified as:

- The number of males in the data set is 2.
- The number of females in the data set is 3.
- The number of individuals identifying as other is 1.
- The number of non-males is 4.

**Univariate vs. Bivariate**

In descriptive statistics, **univariate** data analyzes only one variable. It is used to identify characteristics of a single trait and is not used to analyze any relationships or causations.

For example, imagine a room full of high school students. Say you wanted to gather the average age of the individuals in the room. This univariate data is only dependent on one factor: each person's age. By gathering this one piece of information from each person and dividing by the total number of people, you can determine the average age.

**Bivariate** data, on the other hand, attempts to link two variables by searching for correlation. Two types of data are collected, and the relationship between the two pieces of information is analyzed together. Because multiple variables are analyzed, this approach may also be referred to as multivariate.

Let's say each high school student in the example above takes a college assessment test, and we want to see whether older students are testing better than younger students. In addition to gathering the age of the students, we need to gather each student's test score. Then, using data analytics, we mathematically or graphically depict whether there is a relationship between student age and test scores.

**What is Data Visualization?**

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. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Advantages

- Easily sharing information.
- Interactively explore opportunities.
- Visualize patterns and relationships.

Disadvantages

- Biased or inaccurate information.
- Correlation doesn’t always mean causation.
- Core messages can get lost in translation.

***General Types of Visualizations:***

- **Chart:** Information presented in a tabular, graphical form with data displayed along two axes. Can be in the form of a graph, diagram, or map. Learn more.
- **Table:** A set of figures displayed in rows and columns. Learn more.
- **Graph:** A diagram of points, lines, segments, curves, or areas that represents certain variables in comparison to each other, usually along two axes at a right angle.
- **Geospatial:** A visualization that shows data in map form using different shapes and colors to show the relationship between pieces of data and specific locations. Learn more.
- **Infographic:** A combination of visuals and words that represent data. Usually uses charts or diagrams.
- **Dashboards:** A collection of visualizations and data displayed in one place to help with analyzing and presenting data.

## DATA VISUALIZATION TECHNIQUES

The type of data visualization technique you leverage will vary based on the type of data you're working with, in addition to the story you're telling with your data.

Here are some important data visualization techniques to know:

- Pie Chart
- Bar Chart
- Histogram
- Gantt Chart
- Heat Map
- Box and Whisker Plot
- Waterfall Chart
- Area Chart
- Scatter Plot
- Pictogram Chart
- Timeline
- Highlight Table
- Bullet Graph
- Choropleth Map
- Word Cloud
- Network Diagram
- Correlation Matrices

**R.G. KEDIA COLLEGE OF COMMERCE**
**DEPARTMENT OF BUSINESS MANAGEMENT**
**SUBJECT SYNOPSIS**

**Subject: Business Analytics** **Paper Code: MB-304-S-II**
**Year /Semester: II/III**

**Unit – III: Predictive Analytics:**
Trend Lines, Regression Analysis – Linear & Multiple, Predictive modeling, forecasting Techniques, Data Mining - Definition, Approaches in Data Mining- Data Exploration & Reduction, Data mining and business intelligence, Data mining for business Classification, Association, Cause Effect Modeling.

**What Is a Trendline?**

Trendlines are easily recognizable lines that traders draw on charts to connect a series of prices together or show some data's best fit. The resulting line is then used to give the trader a good idea of the direction in which an investment's value might move.

A trendline is a line drawn over pivot highs or under pivot lows to show the prevailing direction of price. Trendlines are a visual representation of support and resistance in any time frame. They show direction and speed of price, and also describe patterns during periods of price contraction.

What Are the Different Kinds of Trendlines?

There are a number of different kinds of trendlines. The most common are characterized as linear, logarithmic, polynomial, power, exponential, and moving average.

**What Is a Regression?**

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Also called simple regression or ordinary least squares (OLS), linear regression is the most common form of this technique. Linear regression establishes the linear relationship between two variables based on a line of best fit. Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is zero. Non-linear regression models also exist, but are far more complex.

Regression analysis is a powerful tool for uncovering the associations between variables observed in data, but cannot easily indicate causation. It is used in several contexts in business, finance, and economics. For instance, it is used to help investment managers value assets and understand the relationships between factors such as commodity prices and the stocks of businesses dealing in those commodities.

**Calculating Regression**

Linear regression models often use a least-squares approach to determine the line of best fit. The least-squares technique is determined by minimizing the sum of squares created by a mathematical function. A square is, in turn, determined by squaring the distance between a data point and the regression line or mean value of the data set.

Once this process has been completed (usually done today with software), a regression model is constructed. The general form of each type of regression model is:

Simple linear regression:

$$Y = a + bX + u$$

Multiple linear regression:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u$$

**where:**

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

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

$a$ = The y-intercept

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

$u$ = The regression residual or error term

**How Is Regression Analysis Used in Forecasting**

The ***regression method of forecasting*** involves examining the relationship between two different variables, known as the dependent and independent variables. Suppose that you want to forecast future sales for your firm and you've noticed that sales rise or fall, depending on whether the gross domestic product goes up or down. (The gross domestic product, or GDP, is the sum of all goods and services produced within a nation's borders. In the U.S., it is calculated quarterly by the Commerce Department.)

Your sales, then, would be the dependent variable, because they *"depend"* on the GDP, which is the independent variable. (An independent variable is the variable against which you are measuring something by comparison – your sales in this case.) You would need to figure out how closely these two variables - sales and GDP - are related. If the GDP goes up 2 percent, how much do your sales rise?

**What Is Data Mining?**

Data mining is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information.

Companies use data mining software to learn more about their customers. It can help them to develop more effective marketing strategies, increase sales, and decrease costs. Data mining relies on effective data collection, warehousing, and computer processing.

**How Data Mining Works**

Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It is used in credit risk management, fraud detection, and spam filtering. It also is a market research tool that helps reveal the sentiment or opinions of a given group of people. The data mining process breaks down into four steps:

- Data is collected and loaded into data warehouses on-site or on a cloud service.
- Business analysts, management teams, and information technology professionals access the data and determine how they want to organize it.
- Custom application software sorts and organizes the data.
- The end user presents the data in an easy-to-share format, such as a graph or table.

**Data Mining Techniques**

Data mining uses algorithms and various other techniques to convert large collections of data into useful output. The most popular types of data mining techniques include:

- **Association rules**, also referred to as market basket analysis, search for relationships between variables. This relationship in itself creates additional value within the data set as it strives to link pieces of data. For example, association rules would search a company's sales history to see which products are most commonly purchased together; with this information, stores can plan, promote, and forecast.

- **Classification** uses predefined classes to assign to objects. These classes describe the characteristics of items or represent what the data points have in common with each. This data mining technique allows the underlying data to be more neatly categorized and summarized across similar features or product lines.
- **Clustering** is similar to classification. However, clustering identifies similarities between objects, then groups those items based on what makes them different from other items. While classification may result in groups such as "shampoo," "conditioner," "soap," and "toothpaste," clustering may identify groups such as "hair care" and "dental health."
- **Decision trees** are used to classify or predict an outcome based on a set list of criteria or decisions. A decision tree is used to ask for the input of a series of cascading questions that sort the dataset based on the responses given. Sometimes depicted as a tree-like visual, a decision tree allows for specific direction and user input when drilling deeper into the data.
- **K-Nearest neighbor (KNN)** is an algorithm that classifies data based on its proximity to other data. The basis for KNN is rooted in the assumption that data points that are close to each other are more similar to each other than other bits of data. This non-parametric, supervised technique is used to predict the features of a group based on individual data points.
- **Neural networks** process data through the use of nodes. These nodes are comprised of inputs, weights, and an output. Data is mapped through supervised learning, similar to the ways in which the human brain is interconnected. This model can be programmed to give threshold values to determine a model's accuracy.
- **Predictive analysis** strives to leverage historical information to build graphical or mathematical models to forecast future outcomes. Overlapping with regression analysis, this technique aims at supporting an unknown figure in the future based on current data on hand.

**Applications of Data Mining**

In today's age of information, almost any department, industry, sector, or company can make use of data mining.

- Sales
- Marketing
- Manufacturing
- Fraud Detection
- Human Resources
- Customer Service

**Advantages and Disadvantages of Data Mining**

Pros of Data Mining

- It drives profitability and efficiency
- It can be applied to any type of data and business problem
- It can reveal hidden information and trends

Cons of Data Mining

- Complexity
- Results and benefits are not guaranteed
- It can be expensive

**Data Exploration:**

Data exploration is the first step in data analysis involving the use of data visualization tools and statistical techniques to uncover data set characteristics and initial patterns.

During exploration, raw data is typically reviewed with a combination of manual workflows and automated data-exploration techniques to visually explore data sets, look for similarities, patterns and outliers and to identify the relationships between different variables.

This is also sometimes referred to as exploratory data analysis, which is a statistical technique employed to analyze raw data sets in search of their broad characteristics.

**There are several different data reduction techniques that can be used in data mining, including:**

1. **Data Sampling:** This technique involves selecting a subset of the data to work with, rather than using the entire dataset. This can be useful for reducing the size of a dataset while still preserving the overall trends and patterns in the data.
2. **Dimensionality Reduction:** This technique involves reducing the number of features in the dataset, either by removing features that are not relevant or by combining multiple features into a single feature.
3. **Data Compression:** This technique involves using techniques such as lossy or lossless compression to reduce the size of a dataset.
4. **Data Discretization:** This technique involves converting continuous data into discrete data by partitioning the range of possible values into intervals or bins.
5. **Feature Selection:** This technique involves selecting a subset of features from the dataset that are most relevant to the task at hand.
6. It’s important to note that data reduction can have a trade-off between the accuracy and the size of the data. The more data is reduced, the less accurate the model will be and the less generalizable it will be.

In conclusion, data reduction is an important step in data mining, as it can help to improve the efficiency and performance of machine learning algorithms by reducing the size of the dataset. However, it is important to be aware of the trade-off between the size and accuracy of the data, and carefully assess the risks and benefits before implementing it.

**Methods of data reduction:**
These are explained as following below.
**1. Data Cube Aggregation:**
This technique is used to aggregate data in a simpler form. For example, imagine the information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months. They involve you in the annual sales, rather than the quarterly average, So we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.
**2. Dimension reduction:**
Whenever we come across any data which is weakly important, then we use the attribute required for our analysis. It reduces data size as it eliminates outdated or redundant features.
**3. Data Compression:**
The data compression technique reduces the size of the files using different encoding mechanisms (Huffman Encoding & run-length Encoding). We can divide it into two types based on their compression techniques.

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

**4. Numerosity Reduction:**
In this reduction technique, the actual data is replaced with mathematical models or smaller representations of the data instead of actual data, it is important to only store the model parameter. Or non-parametric methods such as clustering, histogram, and sampling.
**5. Discretization & Concept Hierarchy Operation:**
Techniques of data discretization are used to divide the attributes of the continuous nature into data with intervals. We replace many constant values of the attributes by labels of small intervals. This means that mining results are shown in a concise, and easily understandable way.

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

combination of the neighborhood values in the interval, that process is called bottom-up discretization.

**Difference between Business Intelligence and Data Mining**

| Business Intelligence | Data Mining |
|---|---|
| Changing over raw information into valuable data for business. | Designed to investigate information and discover the arrangement for an issue in business. |
| Data-driven makes a difference in choice-making for a business. | Finds answers to an issue or an issue in trade. |
| Expansive Datasets are spokeprocessed on dimensional/social databases Small datasets | handled on a little parcelsKPIsthe of data. |
| Volumetric in nature and display the exact result utilizing visualizations. | Uses calculations to distinguish precise designs for an issue and distinguishes the daze spots. |
| Dashboards and Reports spoken to by charts and charts with KPI's. | Identifies the arrangement for an issue to be spoken to as one of the KPI's in Dashboards or reports. |
| Depending on small-scale past information, there's no intelligence involved; the administration needs to take a choice based on the information. | Focused on a specific issue in trade on small-scale information utilizing calculations to discover the arrangement. |
| Appears cost esteem, benefit, add up to fetch, etc. as KPIs. | Identifies arrangement for an issue making modern KPI's for BI |
| Business Intelligence makes a difference in Decision-making. | Data Mining will unravel a specific issue and contribute to decision-making. |
| Business Intelligence consists of the creation, aggregation, analysis and visualization of data. | Data Mining consists of cleaning, combining, transforming and interpreting of data. |
| | |

**Classification of Data Mining Systems**

Data mining refers to the process of extracting important data from raw data. It analyses the data patterns in huge sets of data with the help of several software. Ever since the development of data mining, it is being incorporated by researchers in the research and development field.

With Data mining, businesses are found to gain more profit. It has not only helped in understanding customer demand but also in developing effective strategies to enforce overall business turnover. It has helped in determining business objectives for making clear decisions.

Data collection and data warehousing, and computer processing are some of the strongest pillars of data mining. Data mining utilizes the concept of mathematical algorithms to segment the data and assess the possibility of occurrence of future events.

To understand the system and meet the desired requirements, data mining can be classified into the following systems:

- Classification based on the mined Databases
- Classification based on the type of mined knowledge
- Classification based on statistics
- Classification based on Machine Learning
- Classification based on visualization
- Classification based on Information Science
- Classification based on utilized techniques
- Classification based on adapted applications

**Cause and Effect Modelling:**

A cause and effect modeling is done to uncover patterns in data of the business organizations.

There are many different types of causal patterns in the world. Below are six patterns that are embedded in many concepts. Causality in the real world seldom falls into one neat pattern or another. The patterns often work together or different parts of a system entail different patterns—making the causality even more complex!

- Linear Causality – Cause precedes effect; sequential pattern. Direct link between cause and effect. Has a clear beginning and a clear ending. Effect can be traced back to one cause. One cause and one effect; additional causes or effects turn this pattern into domino causality
- Domino Causality – Sequential unfolding of effects over time. An extended linear pattern that results in direct and indirect effects. Typically has a clear beginning and a clear

ending. Can be branching where there is more than one effect of a cause (and these may go on to have multiple effects and so on.). Branching forms can be traced back to "stem" causes. Anticipating outcomes involves deciding how far to trace effects. Short-sightedness can lead to unintended effects.

- Cyclic Causality – One thing impacts another which in turn impacts the first thing (or alternatively impacts something else which then impacts something else and so on, but eventually impacts the first thing). Involves a repeating pattern. Involves feedback loops. May be sequential or may be simultaneous. Typically no clear beginning or ending (Sometimes you can look back in time to a beginning but often that results in the classic 'which came first, the chicken or the egg' problem.).
- Spiraling Causality – One thing impacts another which in turn impacts the first thing (or alternatively impacts something else which then impacts something else and so on, but eventually impacts the first thing) with amplification or de-amplification of effects. Involves feedback loops. It is sequential as each event is a reaction to the one before it. Often a clear beginning and ending. It is difficult to anticipate outcomes of later feedback loops during earlier feedback loops.
- Relational Causality – Two things work in relation to each other to cause an outcome. It often involves two variables in comparison to each other. There may be a relationship of balance, equivalence, similarity or there may be a relationship of difference. If one thing changes, so does the relationship, therefore so does the outcome. If two things change but keep the same relationship, the outcome doesn't change.
- Mutual Causality – Two things impact each other. The impact can be positive for both, negative for both, or positive for one and negative for the other. The causes and effects are often simultaneous, but can be sequential. May be event-based or may be a relationship over time

**R.G. KEDIA COLLEGE OF COMMERCE**
**DEPARTMENT OF BUSINESS MANAGEMENT**
**SUBJECT SYNOPSIS**

**Subject: Business Analytics** **Paper Code: MB-304-S-II**
**Year /Semester: II/III**

**Unit – IV: Prescriptive Analytics:**
Overview of Linear Optimization, Non Linear Programming Integer Optimization, Cutting Plane algorithm and other methods, Decision Analysis – Risk and uncertainty methods - Text analytics Web analytics.

### Introduction to Optimization

All optimization models have several common elements:

- **Decision variables** the variable whose values the decision maker is allowed to choose. The values of these variables determine such outputs as total cost, revenue, and profit.
- An **objective function** (**objective**, for short) to be optimized – minimized or maximized.
- **Constraints** that must be satisfied. They are usually physical, logical, or economic restrictions, depending on the nature of the problem.
    - In searching for the values of the decision variables that optimize the objective, only those values that satisfy all of the constraints are allowed.

To **optimize** means that you must systematically choose the values of the decision variables that make the objective as large (for maximization) or small (for minimization) as possible and cause all of the constraints to be satisfied.

Any set of values of the decision variables that satisfies all of the constraints is called a **feasible solution**. The set of all feasible solutions is called the **feasible region**.

### Nonlinear Programming

In many optimization models the objective and/or the constraints are nonlinear functions of the decision variables. Such an optimization model is called a nonlinear programming (NLP) model.

- When you solve an LP model, you are mostly guaranteed that the solution obtained is an optimal solution and a sensitivity analysis with shadow price and reduced cost is available.
- When you solve an NLP model, it is very possible that you obtain a suboptimal solution.

This is because a nonlinear function can have local optimal solutions where

- a local optimum is better than all nearby points

that are not the global optimal solution

- a global optimum is the best point in the entire feasible region

**Cutting Plane Algorithm**:

The cutting plane method is commonly used for solving ILP and MILP problems to find integer solutions, by solving the linear relaxation of the given integer programming model, which is a noninteger LP model. A cutting plane algorithm is generally used to search for valid inequalities that cut-off the noninteger solutions in two cases, when the set of constraints in our integer programming model is too large, and when the inequality constraints in the original integer programming model are not sufficient to yield an integer solution. The basic idea of the cutting plane method is iteratively refining the search region by introducing linear inequalities, known as cuts, maintaining the original feasible region. If the mathematical model is linear, an extreme or corner point in the feasible region can be the optimal solution, which may or may not be integer solution. If it is not integer, a linear inequality, in other words a cut, can be found to cut away a part of the feasible region, so as to separate the optimum solution. Therefore, a new separation problem is introduced and a cut is added to the relaxed LP model, which makes the existing noninteger solution no longer in the feasible region. This cutting process is repeated until the optimal solution found is also an integer solution.
The steps of the cutting plane algorithm can be explained as follows:

1. The relaxed integer programming problem (the problem with continuous variables instead of discrete/integer variables) is solved.
2. Stop, if all variables in resulting solution have integer values, which means that it is the optimum.
3. Otherwise, generate a cut, that is, a constraint in the form of a linear inequality, which is satisfied by all feasible integer solutions.
4. Add this new constraint to the model, resolve the problem, and go back to step 2.

**What Is Decision Analysis (DA)?**

Decision analysis (DA) is a systematic, quantitative, and visual approach to addressing and evaluating the important choices that businesses sometimes face. Ronald A. Howard, a professor of Management Science and Engineering at Stanford University, is credited with originating the term in 1964.1 The idea is used by large and small corporations alike when making various types of decisions, including management, operations, marketing, capital investments, or strategic choices.

## How to Make Decisions Under Uncertainty?

Decision-making under uncertainty involves situations where the outcome of a decision is unknown. Decision-makers must consider multiple possible outcomes and their probabilities in such cases. There are several techniques that decision-makers can use to make decisions under uncertainty, including the Laplace criterion, Maximin, Maximax, Hurwicz, and Minimax regret.

### Laplace Criterion

The Laplace criterion is a decision-making technique that can be utilized to make decisions under uncertainty using AI. It is used to make decisions in situations where the probability of each outcome is unknown or cannot be estimated. This technique assumes that each outcome is equally likely and assigns equal weight to each of them. In this method, we calculate the average value of each scenario irrespective of the probability of occurrence of each scenario. The decision-makers select the scenario with the highest average value. AI tools support decision-makers in putting the Laplace criterion into practice. Large datasets are analyzed, and the average result for each decision is computed. AI can be used to find trends and patterns in data. The future results of any decision can be predicted with the assistance of these patterns.

### Maximin

The Maximin criterion is a decision-making technique that can be used to make decisions under uncertainty using AI. When faced with a situation where the probability of each outcome is unknown or cannot be estimated, decision-makers can employ this technique to select the best course of action. The Maximin criterion assumes that the worst possible outcome of a decision is the most important consideration. It means that decision-makers must consider the most negative result of each decision.

AI tools support decision-makers in putting the Maximin criterion into practice. Large datasets are analyzed, and the worst possible outcome for each decision is computed. AI can be used to find trends and patterns in data. We can predict the future results of any decision with the assistance of these patterns. It is worth noting that the Maximin criterion may not always lead to the optimal decision since it only considers the worst possible outcome. Nonetheless, it can be an effective technique when decision-makers prioritize avoiding the worst-case scenario.

Let's say you are playing a game with a friend and have to choose between two options: Option A gives you a guaranteed win of Rs. 5/-, while Option B provides a 50-50 chance of winning either Rs.10/- or Rs.0/-. If you use the Maximin strategy, you will choose Option A because it gives you the highest minimum payoff (i.e., Rs. 5/-) compared to Option B, which has a minimum payoff of Rs. 0/- if you lose the coin toss.

### Maximax

The Maximax criterion is a decision-making technique that can be used to make decisions under uncertainty using AI. When faced with a situation where the probability of each outcome is unknown or cannot be estimated, decision-makers can employ this technique to select the best course of action. The Maximax criterion assumes that the best possible outcome of a decision is the most important consideration. The decision with the highest outcome is selected as the output.

Suppose you are a product lead manager considering launching a new product. You have estimated the potential profits for the new product under three different scenarios: low demand, moderate demand, and high demand. The estimated profits for each scenario are as follows: Low demand will give Rs. 10,000/- profit, moderate demand would provide Rs. 50,000/- profit, and a high demand would give Rs. 1,00,000/- profit. Using the Maximax strategy, you would choose the option that maximizes the potential profit in the best-case scenario. The best-case scenario is high demand, potentially generating Rs. 1,00,000/- profit. Therefore, you should launch the new product because it has the highest potential profit in the best-case scenario.

**Hurwicz**

The Hurwicz technique helps choose a decision that balances good and bad outcomes. This technique uses a coefficient that decides how much weight should be given to the best and worst outcomes. The best and worst outcomes are equally important when the coefficient's value equals 0.5. Only the best outcome is considered if the coefficient takes a value equal to 1. On the other hand, only the worst outcome is considered if the coefficient takes a value equal to 0.

**Minimax Regret**

The Minimax regret technique involves choosing the decision that minimizes the maximum regret. Regret is the difference between the best outcome and the outcome of the chosen decision. This technique aims to minimize the regret of making a suboptimal decision.

**How to Make Decisions Under Risk?**

Decision-making under risk involves situations where the probability of each outcome is known or can be estimated. This helps decision-makers to use statistical methods to analyze the options and make the best decision. To make decisions under risk, decision-makers might employ a variety of strategies as discussed below:

**Maximum Expected Value**

The maximum expected value method is the first way that decision-makers might use to assess and contrast various options. Finding the option with the highest expected value is the key to employing this strategy. The expected value is calculated as the product of all possibilities and their corresponding probabilities. The choice that has the best probability of rewarding them or offering benefits can subsequently be made by decision-makers. This method is beneficial when decision-makers are judged to be risk-neutral, that is, neither risk-avoiders nor risk-seeking, and are therefore only concerned with maximizing their expected gains. Using the maximum expected value technique, decision-makers can decide the optimal course of action to assist them in reaching their intended goals by basing their decisions on the probability associated with each possible outcome.

Suppose you are a sales manager and must decide whether to launch a new product line. You have estimated the potential profits for the new product line under two different scenarios:
a 90% chance of moderate demand, which would result in an Rs. 50,000/- profit, and
a 10% chance of low demand, which would result in an Rs. 10,000/- loss. To use the Maximum Expected Value strategy, you would multiply each scenario's potential profit or loss by its probability and add them together. In this case, the expected value of launching the new product line would be:

Expected Value = (0.90 x Rs. 50,000) + (0.10 x Rs. 10,000)
Expected Value = Rs. 46,000/-

The expected value of launching the new product line is Rs. 46,000/-, which is the sum of the potential profits and losses weighted by their probabilities. Since the expected value is positive, you can launch the new product line since it has the highest expected profit. However, it is important to note that the Maximum Expected Value strategy does not consider each scenario's potential risks or uncertainties.

**Maximum Utility**

The maximum utility technique is a popular approach to decision-making that involves selecting the decision with the highest expected utility. Utility refers to the value or satisfaction a decision can provide an individual. In contrast to the maximum expected value technique, the maximum utility technique considers risk-avoiding decision-makers and assigns more significant value to outcomes with a lower probability. This technique is often used when decision-makers want to maximize the overall satisfaction of a decision rather than merely its expected gains. This technique allows decision-makers to account for the potential trade-offs between risks and rewards, which can help them make more informed decisions. Additionally, the maximum utility technique can help decision-makers to prioritize their preferences and determine which outcomes are most desirable, given their risk tolerance and personal preferences. Overall, the maximum utility technique is a powerful tool for decision-makers who want to optimize their decision-making process under conditions of uncertainty and risk.

**Most Probable Outcome**

Decision-makers may opt to use the most probable outcome technique when a risky option is present. This approach decides with a high chance of success. This method seeks to reduce the level of uncertainty while making decisions. It helps in giving a thorough understanding of the possible outcomes. This approach can be especially useful when decision-makers seek to avoid risks and want some degree of certainty in their choice. By selecting the most probable outcome, decision-makers can make well-informed decisions with high confidence.

**Composite Criteria**

The composite criteria technique includes integrating several factors to produce a judgment. This method gives several criteria with varied weights while considering the preferences and priorities of the decision-maker.

**Analysis Methods**

Decision-makers can examine the results of their decisions in a variety of ways. These techniques include Monte Carlo simulations, sensitivity analysis, and decision trees. Decision trees visually represent a decision's potential outcomes and associated probabilities. They help figure out the best course of action and visualize complicated decisions. Sensitivity analysis comprises altering the inputs and assumptions used in the study to see how robust a choice is. This method can assist decision-makers in locating the crucial elements that influence a decision's outcome. In Monte Carlo simulations, the variables employed in the analysis are given random values, and the potential outcomes of a choice are simulated. With the help of this technique, decision-makers can better comprehend the range of potential outcomes and the likelihood of each one.

**What is Text Analytics?**

**Text analytics** is the process of transforming unstructured text documents into usable, structured data. Text analysis works by breaking apart sentences and phrases into their components, and then evaluating each part's role and meaning using complex software rules and machine learning algorithms.

Text analytics forms the foundation of numerous natural language processing (NLP) features, including named entity recognition, categorization, and sentiment analysis. In broad terms, these NLP features aim to answer four questions:

- Who is talking?
- What are they talking about?
- What are they saying about those subjects?
- How do they feel?

Data analysts and other professionals use text mining tools to derive useful information and context-rich insights from large volumes of raw text, such as social media comments, online reviews, and news articles. In this way, text analytics software forms the backbone of business intelligence programs, including voice of customer/customer experience management, social listening and media monitoring, and voice of employee/workforce analytics.

**Web Analytics:**

Web analytics is the gathering, synthesizing, and analysis of website data with the goal of improving the website user experience. It's a practice that's useful for managing and optimizing websites, web applications, or other web products. It's highly data-driven and assists in making high-quality website decisions. You might also get ideas on how to improve your product and drive business growth from web analytics.

**What is web analytics used for?**

Web analytics is helpful for understanding which channels users come through to your website. You can also identify popular site content by calculating the average length of stay on your web pages and how users interact with them—including which pages prompt users to leave.

The process of web analytics involves:

- **Setting business goals:** Defining the key metrics that will determine the success of your business and website
- **Collecting data:** Gathering information, statistics, and data on website visitors using analytics tools
- **Processing data:** Converting the raw data you've gathered into meaningful ratios, KPIs, and other information that tell a story
- **Reporting data**: Displaying the processed data in an easy-to-read format
- **Developing an online strategy:** Creating a plan to optimize the website experience to meet business goals
- **Experimenting**: Doing **A/B tests** to determine the best way to optimize website performance

## Unit – V: Programming Using R:

R Environment, R packages, Reading and Writing data in R, R functions, Control Statements, Frames and Subsets, Managing and Manipulating data in R.

### Environments in R Programming

The environment is a virtual space that is triggered when an interpreter of a programming language is launched. Simply, the environment is a collection of all the objects, variables, and functions. Or, Environment can be assumed as a top-level object that contains the set of names/variables associated with some values.

Why the Environment Differ from the List?

- Every object in an environment has a name.
- The environment has a parent environment.
- Environments follow reference semantics.

***Create a New Environment***

An environment in R programming can be created using **new.env()** function. Further, the variables can be accessed using **$ or [[ ]]** operator. But, each variable is stored in different memory locations. There are four special environments: **globalenv(), baseenv(), emptyenv() and environment()**

***Syntax:*** *new.env(hash = TRUE)*

***Parameters:***

***hash:*** *indicates logical value. If TRUE, environments uses a hash table*

### R - Packages

R packages are a collection of R functions, complied code and sample data. They are stored under a directory called **"library"** in the R environment. By default, R installs a set of packages during installation. More packages are added later, when they are needed for some specific purpose. When we start the R console, only the default packages are available by default. Other packages which are already installed have to be loaded explicitly to be used by the R program that is going to use them.

All the packages available in R language are listed at R Packages.

Packages in library ‘C:/Program Files/R/R-3.2.2/library’:

| | |
|---|---|
| base | The R Base Package |
| boot | Bootstrap Functions (Originally by Angelo Canty for S) |
| class | Functions for Classification |
| cluster | "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. |
| codetools | Code Analysis Tools for R |
| compiler | The R Compiler Package |
| datasets | The R Datasets Package |
| foreign | Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... |
| graphics | The R Graphics Package |
| grDevices | The R Graphics Devices and Support for Colours and Fonts |
| grid | The Grid Graphics Package |
| KernSmooth | Functions for Kernel Smoothing Supporting Wand & Jones (1995) |
| lattice | Trellis Graphics for R |
| MASS | Support Functions and Datasets for Venables and Ripley's MASS |
| Matrix | Sparse and Dense Matrix Classes and Methods |
| methods | Formal Methods and Classes |
| mgcv | Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation |
| nlme | Linear and Nonlinear Mixed Effects Models |
| nnet | Feed-Forward Neural Networks and Multinomial Log-Linear Models |
| parallel | Support for Parallel computation in R |
| rpart | Recursive Partitioning and Regression Trees |
| spatial | Functions for Kriging and Point Pattern Analysis |
| splines | Regression Spline Functions and Classes |
| stats | The R Stats Package |
| stats4 | Statistical Functions using S4 Classes |

| | |
|---|---|
| survival | Survival Analysis |
| tcltk | Tcl/Tk Interface |
| tools | Tools for Package Development |
| utils | The R Utils Package |

**R – Writing to Files**

**Writing Data to CSV files in R Programming Language**

CSV stands for Comma Separated Values. These files are used to handle a large amount of statistical data. Following is the syntax to write to a CSV file:

**Syntax:**

- R

```
write.csv(my_data, file = "my_data.csv")

write.csv2(my_data, file = "my_data.csv")
```

*csv() and csv2() are the function in R programming.*

- ***write.csv()*** *uses "." for the decimal point and a comma (", ") for the separator.*
- ***write.csv2()*** *uses a comma (", ") for the decimal point and a semicolon (";") for the separator.*

**Writing Data to text files**

Text files are commonly used in almost every application in our day-to-day life as a step for the "Paperless World". Well, writing to .txt files is very similar to that of the CSV files. Following is the syntax to write to a text file:

**Syntax:**

- R

```
write.table(my_data, file = "my_data.txt", sep = "")
```

**R - Functions**

A function is a set of statements organized together to perform a specific task. R has a large number of in-built functions and the user can create their own functions.

In R, a function is an object so the R interpreter is able to pass control to the function, along with arguments that may be necessary for the function to accomplish the actions.

The function in turn performs its task and returns control to the interpreter as well as any result which may be stored in other objects.

Function Definition

An R function is created by using the keyword **function**. The basic syntax of an R function definition is as follows −

```
function_name <- function(arg_1, arg_2, ...) {
  Function body
}
```

Function Components

The different parts of a function are −

- **Function Name** − This is the actual name of the function. It is stored in R environment as an object with this name.
- **Arguments** − An argument is a placeholder. When a function is invoked, you pass a value to the argument. Arguments are optional; that is, a function may contain no arguments. Also arguments can have default values.
- **Function Body** − The function body contains a collection of statements that defines what the function does.
- **Return Value** − The return value of a function is the last expression in the function body to be evaluated.

**Control Statements in R Programming**

**Control statements** are expressions used to control the execution and flow of the program based on the conditions provided in the statements. These structures are used to make a decision after assessing the variable. In this article, we'll discuss all the control statements with the examples.

In R programming, there are 8 types of control statements as follows:

- if condition
- if-else condition
- for loop
- nested loops
- while loop

- repeat and break statement
- return statement
- next statement

***if condition***

This control structure checks the expression provided in parenthesis is true or not. If true, the execution of the statements in braces {} continues.

**Syntax:**

```
if(expression){
   statements
   ....
   ....
}
```

***if-else condition***

It is similar to **if** condition but when the test expression in if condition fails, then statements in **else** condition are executed.

**Syntax:**

```
if(expression){
   statements
   ....
   ....
}
else{
   statements
   ....
   ....
}
```

***for loop***

It is a type of loop or sequence of statements executed repeatedly until exit condition is reached.

**Syntax:**

```
for(value in vector){
   statements
   ....
   ....
```

}

***while loop***

while loop is another kind of loop iterated until a condition is satisfied. The testing expression is checked first before executing the body of loop.

**Syntax:**

```
while(expression){
    statement
    ....
    ....
}
```

***repeat loop and break statement***

**repeat** is a loop which can be iterated many number of times but there is no exit condition to come out from the loop. So, break statement is used to exit from the loop. **break** statement can be used in any type of loop to exit from the loop.

**Syntax:**

```
repeat {
   statements
   ....
   ....
   if(expression) {
      break
   }
}
```

***return statement***

**return** statement is used to return the result of an executed function and returns control to the calling function.

**Syntax:**

return(expression)

**Data Manipulation in R**

In order to manipulate the data, R provides a library called dplyr which consists of many built-in methods to manipulate the data. So to use the data manipulation function, first need to import the dplyr package using ***library(dplyr)*** line of code. Below is the list of a few data manipulation functions present in dplyr package.

| Function Name | Description |
|---|---|
| filter() | Produces a subset of a Data Frame. |
| distinct() | Removes duplicate rows in a Data Frame |
| arrange() | Reorder the rows of a Data Frame |
| select() | Produces data in required columns of a Data Frame |
| rename() | Renames the variable names |
| mutate() | Creates new variables without dropping old ones. |
| transmute() | Creates new variables by dropping the old. |
| summarize() | Gives summarized data like Average, Sum, etc. |