MARUDHAR KESARI JAIN COLLEGE FOR WOMEN (AUTONOMOUS)

VANIYAMBADI

PG Department of Computer Applications

I MCA – Semester - II

E-Notes (Study Material)

Core Course -1: Data Analytics and Visualization	**Code: 24PCAC21**
Unit: 2 - Introduction to Analytics, Introduction to Tools and Environment, Application of Modeling in Business, Databases & Types of Data and variables, Data Modeling Techniques, Missing Imputation setc. Need for Business Modeling.	
Learning Objectives: To learn the principles and methods of statistical analysis	
Course Outcome: To understand the principles and methods of statistical analysis	

Overview:

- Introduction to Analytics
- Tools & Environment
- Application and Types of Data
-Data Modelling Techinques
- Need for Business Model

1. Introduction to Analytics:

Analytics is a journey that involves a combination of potential skills, advanced technologies, applications, and processes used by firm to gain business insights from data and statistics. This is done to perform business planning.

Data Analytics refers to the techniques used to analyze data to enhance productivity and business gain. Data is extracted from various sources and is cleaned and categorized to analyze various behavioral patterns. The techniques and the tools used vary according to the organization or individual.

Data Analytics has a key role in improving your business as it is used to gather hidden insights, generate reports, perform market analysis, and improve business requirements.

Data analytics is the process of inspecting, transforming and Extract Meaningful Insights from data for Decision making

Data analytics is a scientific process of Convert Data into Useful Information for Decision Makers

Role of Data Analytics:

Gather Hidden Insights – Hidden insights from data are gathered and then analyzed with respect to business requirements.

Generate Reports – Reports are generated from the data and are passed on to the respective teams and individuals to deal with further actions for a high rise in business.

Perform Market Analysis – Market Analysis can be performed to understand the strengths and weaknesses of competitors.

Improve Business Requirement – Analysis of Data allows improving Business to customer requirements and experience.

Applications of Analytics:

In today's world, data rules the most modern companies.To gain such important insight into data as a whole, it is important to analyze data and draw specific information that can be used to improve certain aspects of a market or the business as a whole.

There are several applications of data analytics, and businesses are actively using such data analytics applications to keep themselves in the competition.

- **Fraud and Risk Detection:**

This has been known as one of the initial applications of data science which was extracted from the discipline of Finance. So many organizations had very bad experiences with debt and were so fed up with it.

Since they already had data that was collected during the time their customers applied for loans, they applied data science which eventually rescued them from the losses they had incurred.

This led to banks learning to divide and conquer data from their customers' profiles, recent expenditure and other significant information that were made available to them.

This made it easy for them to analyze and infer if there was any probability of customers defaulting.

- **Policing/Security**

Several cities all over the world have employed predictive analysis in predicting areas that would likely witness a surge in crime with the use of geographical data and historical data.

This has seemed to work in major cities such as Chicago, London, Los Angeles, etc. Although, it is not possible to make arrests for every crime committed but the availability of data has made it possible to have police officers within such areas at a certain time of the day which has led to a drop in crime rate.

This shows that this kind of data analytics application will make us have safer cities without police putting their lives at risk.

- **Healthcare:**

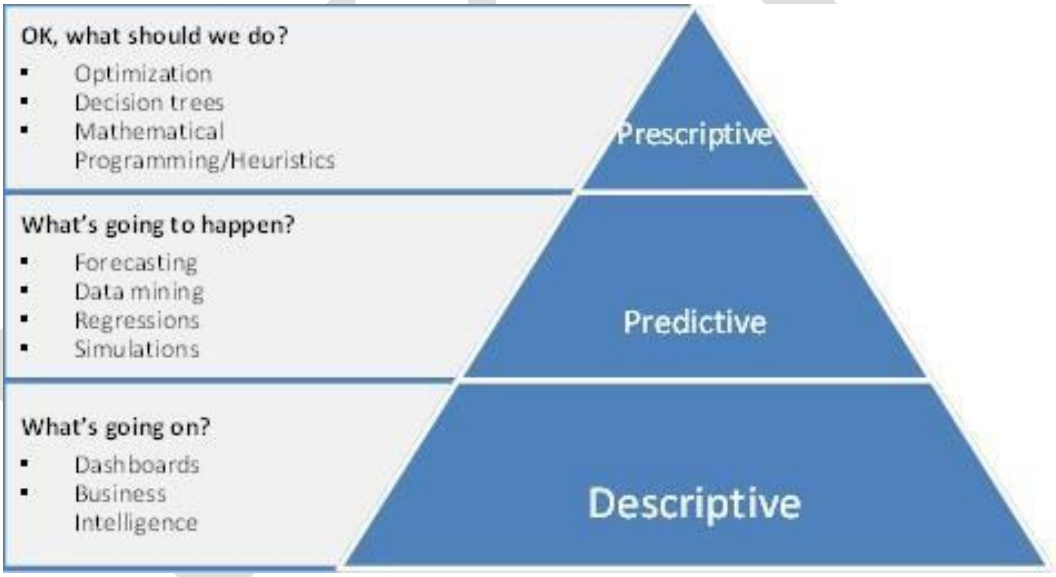
Healthcare analytics is the process of analyzing current and historical industry data to predict trends, improve outreach, and even better manage the spread of diseases. The field covers a broad range of businesses and offers insights on both the macro and micro level.

It can reveal paths to improvement in patient care quality, clinical data, diagnosis, and business management.

When combined with business intelligence suites and data visualization tools, healthcare analytics help managers operate better by providing real-time information that can support decisions and deliver actionable insights.

Types of data analytics:

There are 3 different types of analytics. Here, we start with the simplest one and go further to the more sophisticated types. As it happens, the more complex an analysis is, the more value it brings.



I) Descriptive Analytics

- It describes what has already occurred. It helps the business to understand how things are going.
- Data aggregation and data mining are two techniques used indescriptive analytics to discover historical data.

- Data is first gathered and sorted by data aggregation in order to make the datasets more manageable by analysts.
- Data mining describes the next step of the analysis and involves a search of the data to identify patterns and meaning.
- Identified patterns are analyzed to discover the specific ways that learners interacted with the learning content and within the learning environment.

2) Predictive Analytics

- It tells us what will probably happen in the future as a result of something that has already happened.
- It takes historical data and feeds it into a machine learning model that considers key trends and patterns. The model is then applied to current data to predict what will happen next.
- Which use statistical models and forecasts techniques to understand the future and answer: "What could happen?"

3) Prescriptive Analytics

- Prescriptive analytics is a statistical method used to generate recommendations and make decisions based on the computational findings of algorithmic models.
- Prescriptive analytics is an emerging discipline and represents a more advanced use of predictive analytics.
- Which use optimization and simulation algorithms to advice on possible outcomes and answer: "What should we do?"
- Google's self-driving caris an example of prescriptive analytics in action. The vehicle makes millions of calculations on every trip that helps the car decide when and where to turn, whether to slow down or speed up and when to change lanes — the same decisions a human driver makes behind the wheel.

Introduction to tools and Environment:

- Analytics is now days used in all the fields ranging from Medical Science to Aero science to Government Activities.
- Data Science and Analytics are used by Manufacturing companies as well as Real Estate firms to develop their business and solve various issues by the help of historical data base.
- Tools are the softwares that can be used for Analytics like SAS or R. While techniques are the procedures to be followed to reach up to a solution.
- Various steps involved in Analytics:
    1. Access
    2. Manage
    3. Analyze
    4. Report

Various Analytics techniques are:

1. Data Preparation
2. Reporting, Dashboards & Visualization
3. Segmentation Icon
4. Forecasting
5. Descriptive Modeling
6. Predictive Modeling
7. Optimization

R Programming:

- R is a statistical language created by statisticians. Thus, it excels in statistical computation. R is the mostused programming language for developing statistical tools.
- R compiles and runs on various platforms such as UNIX, Windows and Mac OS.
- R is helpful at every step of the data analysis process from gathering and cleaning data to analyzing it and reporting the conclusions.
- It can easily manipulate your data and present in different ways.
- R also integrates very well with many Big Data platforms which have contributed to its success.
- R has vast number of packages and built in functions.
- R facilitates quality plotting and graphing. The popular libraries like ggplot2 forvisually appealing graphs that set R apart from other programming languages.

Python:

- Python is a high level, interpreted and general purpose dynamic programming language that focuses on code readability.It was founded in 1991 by developer Guido Van Rossum.It is used in many organizations as it supports multiple programming paradigms.
- It also performs automatic memory management.
- You need less lines of code to perform the same task as compared to other major languages like C/C++ and Java.
- Python is a very productive language. Due to the simplicity of Python, developers can focus on solving the problem.
- Python provides a large standard library which includes areas like internet protocols, string operations, web services tools and operating system interfaces.you can find almost all the functions needed for your task.
- Python has built-in list and dictionary data structures which can be used to construct fast runtime data structures.

Tableau Public:

- Tableau Public is a free software that connects any data source be it corporate Data Warehouse, Microsoft Excel or web-based data, and creates data visualizations, maps, dashboards etc. with real-time updates presenting on web.
- It is the perfect visualization tool used for analysis.
- Most suitable for quick and easy representation of big data which helps in resolving the big data issues.
- Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data into the very easily understandable format.
- Data analysis is very fast with Tableau and the visualizations created are in the form of dashboards and worksheets.

- The data that is created using Tableau can be understood by professional at any level in an organization.
- It even allows a non-technical user to create a customized dashboard.
- The great thing about Tableau software is that it doesn't require any technical or any kind of programming skills to operate.
- The University of California, Berkeley's AMP Lab, developed Apache in 2009. Apache Spark is a fast large-scale data processing engine and executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk.

➢ **Apache Spark** – One of the largest large-scale data processing engine, this tool executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk. This tool is also popular for data pipelines and machine learning model development.

➢ **QlikView** – This tool offers in-memory data processing with the results delivered to the end-users quickly. It also offers data association and data visualization with data being compressed to almost 10% of its original size.

- **SAS** – A programming language and environment for data manipulation and analytics, this tool is easily accessible and can analyze data from different sources.

- **Microsoft Excel** – This tool is one of the most widely used tools for data analytics. Mostly used for clients' internal data, this tool analyzes the tasks that summarize the data with a preview of pivot tables.

- **RapidMiner** – A powerful, integrated platform that can integrate with any data source types such as Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase etc. This tool is mostly used for predictive analytics, such as data mining, text analytics, machine learning.

Types of Data Models:

1. Hierarchical Model
2. Relational Model
3. Network Model
4. Object-Oriented Model
5. Entity-Relationship Model

l. Hierarchical Model

As the name indicates, this model makes use of hierarchy to structure the data in a tree-like format. However, retrieving and accessing data is difficult in hierarchical model.

The hierarchy starts from the root which has root data and then it expands in the form of a tree adding child node to the parent node. This model easily represents some of the real-world relationships like food recipes, sitemap of a website etc.

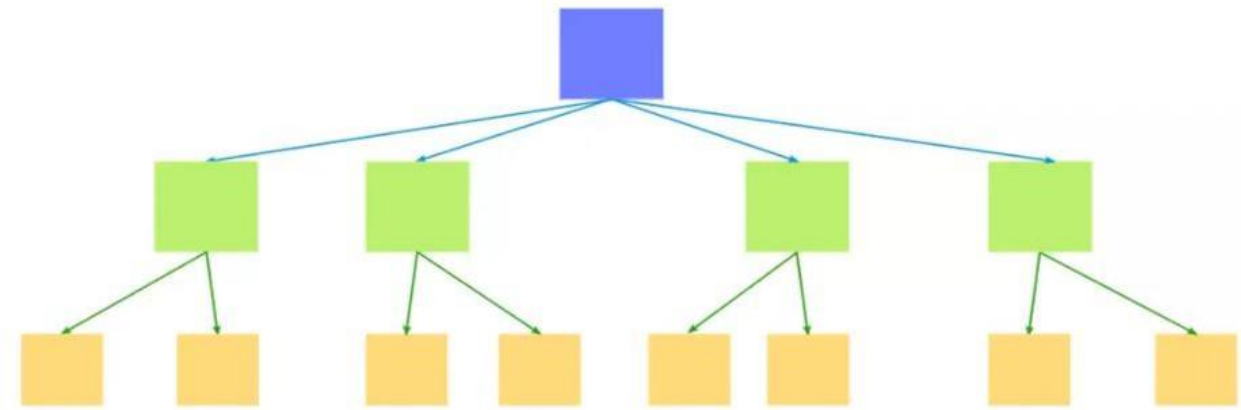
Advantages:

- It is very simple and fast to traverse through a tree-like structure.
- Any change in the parent node is automatically reflected in the child node so, the integrity of data is maintained.

Disadvantages:

- Complex relationships are not supported.

- As it does not support more than one parent of the child node so if we have some complex relationship where a child node needs to have two parent node then that can't be represented using this model.
- If a parent node is deleted then the child node is automatically deleted.

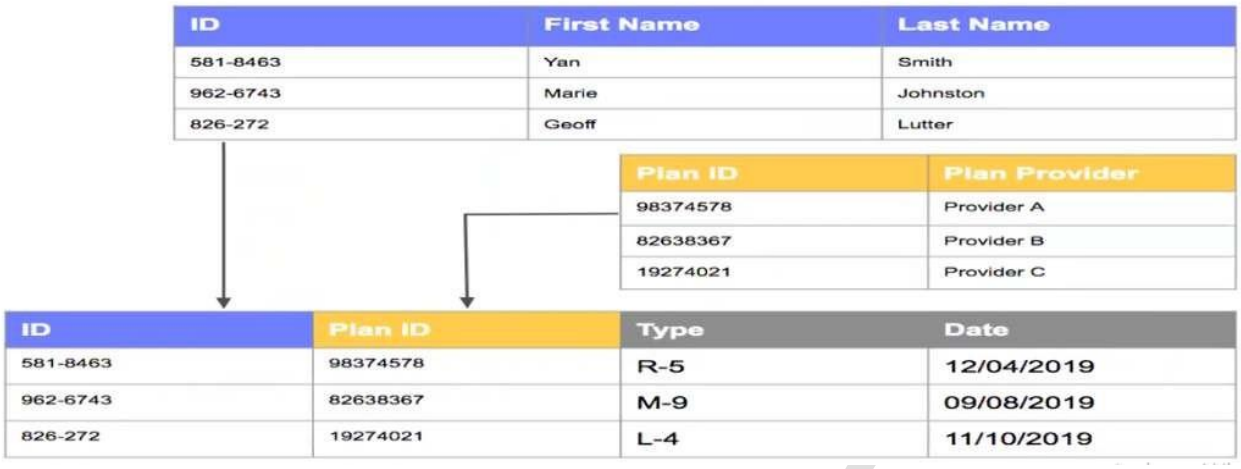


2. Relational Model

Relational Model is the most widely used model. In this model, the data is maintained in the form of a two-dimensional table. All the information is stored in the form of row and columns. The basic structure of a relational model is tables.

Advantages:

- **Simple:** This model is more simple as compared to the network and hierarchical model.
- **Scalable:** This model can be easily scaled as we can add as many rows and columns we want.
- **Structural Independence:** We can make changes in database structure without changing the way to access the data. When we can make changes to the database structure without affecting the capability to DBMS to access the data we can say that structural independence has been achieved.



ID	First Name	Last Name
581-8463	Yan	Smith
962-6743	Marie	Johnston
826-272	Geoff	Lutter

Plan ID	Plan Provider
98374578	Provider A
82638367	Provider B
19274021	Provider C

ID	Plan ID	Type	Date
581-8463	98374578	R-5	12/04/2019
962-6743	82638367	M-9	09/08/2019
826-272	19274021	L-4	11/10/2019

3. Network Model

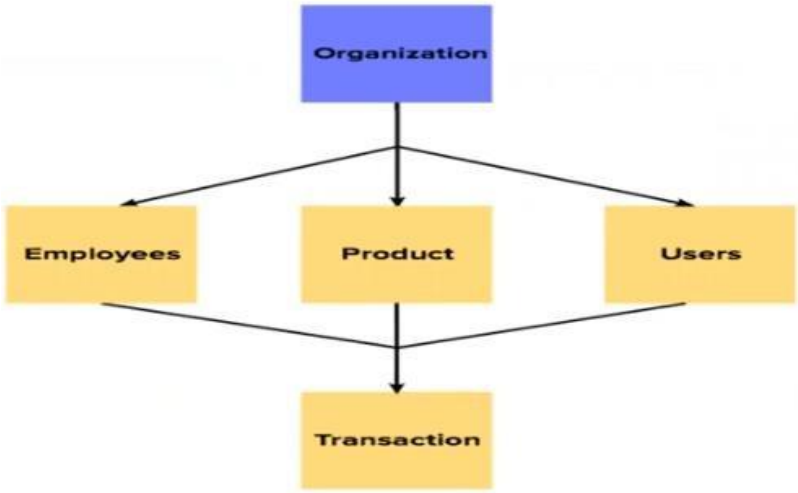
- The network model is an extension of the hierarchical model. However, unlike the hierarchical model, this model makes it easier to convey complex relationships as each record can be linked with multiple parent records.
- The network model is a database model conceived as a flexible way of representing objects and their relationships. Its distinguishing feature is that the schema, viewed as a graph in which object types are nodes and relationship types are arcs, is not restricted to being a hierarchy or lattice.

Advantages:

- The data can be accessed faster as compared to the hierarchical model. This is because the data is more related in the network model and there can be more than one path to reach a particular node. So the data can be accessed in many ways.
- As there is a parent-child relationship so data integrity is present. Any change in parent record is reflected in the child record.

Disadvantages:

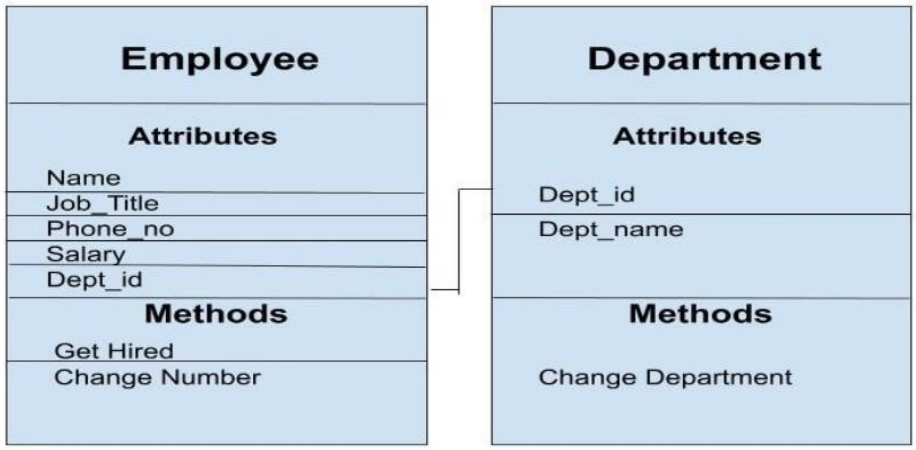
- As more and more relationships need to be handled the system might get complex. So, a user must be having detailed knowledge of the model to work with the model.
- Any change like updation, deletion, insertion is very complex.



4. Object oriented Model:

This model consists of a collection of objects, each with its own features and methods. This type of model is also called the post relational database model.

The real-world problems are more closely represented through the object-oriented data model.
In this model, two are more objects are connected through links. We use this link to relate one object to other objects.



Object_Oriented_Model

In the above example, we have two objects Employee and Department. All the data and relationships of each object are contained as a single unit.

The attributes like Name, Job_title of the employee and the methods which will be performed by that object are stored as a single object. The two objects are connected through a common attribute i.e the Department_id and the communication between these two will be done with the help of this common id.

5. Entity-relationship

Entity-Relationship Model or simply ER Model is a high-level data model diagram.

In this model, we represent the real-world problem in the pictorial form to make it easy for the stakeholders to understand.

An entity could be anything – a concept, a piece of data, or an object.

Advantages:

- ***Simple:*** Conceptually ER Model is very easy to build. If we know the relationship between the attributes and the entities we can easily build the ER Diagram for the model.
- ***Effective Communication Tool***: This model is used widely by the database designers for communicating their ideas.
- ***Easy Conversion to any Model***: This model maps well to the relational model and can be easily converted relational model by converting the ER model to the table. This model can also be converted to any other model like network model, hierarchical model etc.

Disadvantages:

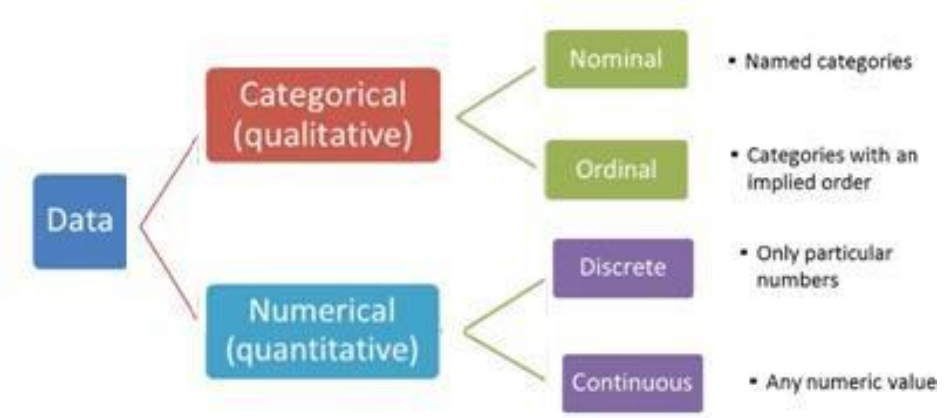
- **No industry standard for notation:** There is no industry standard for developing an ER model. So one developer might use notations which are not understood by other developers.
- **Hidden information:** Some information might be lost or hidden in the ER model. As it is a high-level view so there are chances that some details of information might be hidden

Types of Data:

Data Types are an important concept of statistics, which needs to be understood, to correctly apply statistical measurements to your data and therefore to correctly conclude certain assumptions about it.

There are two types of variables you'll find in your data

- ❖ Numerical(Quantitative)
- ❖ Categorical(Qualitative)



l. Quantitative data (Numerical data):

It deals with numbers and things you can measure objectively: dimensions such as height, width, and length. Temperature and humidity, Prices, Area and volume. Numerical data is information that is measurable, and it is, of course, data represented as numbers and not words or text.

Numerical data can be divided into continuous ordiscrete values.

✓ **Continuous Data:**

Continuous Data represents measurements and therefore their values can't be counted but they can be measured.

Continuous numbers are numbers that don't have a logical end to them.

An example would be the height of a person, which you can describe by using intervals on the real number line.

✓ **Discrete Data:**

We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values.

Discrete numbers are the opposite; they have a logical end to them.

Some examples include variables for days in the month, or number of bugs logged.

2. Categorical Data

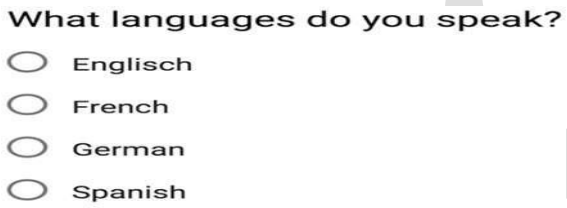
Categorical data represents characteristics. This is any data that isn't a number, which can mean a string of text or date.

Therefore it can represent things like a person's gender, language etc.

These variables can be broken down into nominal and ordinal values.

- **Nominal Data**

Nominal values represent discrete units and are used to label variables, that have no quantitative value. Nominal value examples include variables such as "Country" or "Marital Status".



- **Ordinal data**

Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters.

Examples of ordinal values include having a priority on a bug such as "Critical" or "Low" or the ranking of a race as "First" or "Third".

- **Binary data**

In addition to ordinal and nominal values, there is a special type of categorical data called binary. Binary data types only have two values – **yes or no**.

This can be represented in different ways such as "True" and "False" or 1 and 0.

Examples of binary variables can include whether a person has stopped their subscription service or not, or if a person bought a car or not.

Missing Imputations:

In R, missing values are represented by the symbol **NA** (not available). Impossible values (e.g.,dividing by zero) are represented by the symbol **NaN**(not a number). To remove missing values from our dataset we use ***na.omit()*** function.

For Example:

We can create new dataset without missing data as below:

newdata<-na.omit(mydata)

Or,

we can also use "**na.rm=TRUE**" in argument of the operator. From above example we use na.rm and get desired result.

```
x <- c(1,2,NA,3)
mean(x, na.rm=TRUE)
# returns 2
```

Missing Imputations (MICE Package)

MICE : MICE (Multivariate Imputation via Chained Equations) is one of the commonly used package by R users. Creating multiple imputations as compared to a single imputation takes care of uncertainty in missing values.

The mice package implements a method to deal with missing data. The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data.

For example:

Suppose we have X1, X2….Xk variables. If X1 has missing values, then it will be regressed on other variables X2 to Xk. The missing values in X1 will be then replaced by predictive values obtained. Similarly, if X2 has missing values, then X1, X3 to Xk variables will be used in prediction model as independent variables. Later, missing values will be replaced with predicted values. mice package has a function known as ***md.pattern().*** It returns a tabular form of missing value present in each variable in a data set.

Syntax:

imputed_Data<- mice(data, m=5, maxit = 5, method = 'NULL', seed = NA)

Precisely, the methods used by this package are:

1. **PMM** (Predictive Mean Matching)
   – For numeric variables.
2. **logreg**(Logistic Regression)
   – For Binary Variables(with 2 levels)
3. **polyreg**
   – For Factor Variables (>= 2 levels)
4. **Proportional odds model**
   (ordered, >= 2 levels)

Application of Modeling in Business:

A statistical model embodies a set of assumptions concerning the generation of the observed data, and similar data from a larger population. A model represents, often in considerably idealized form, the data-generating process.

Signal processing is an enabling technology that encompasses the fundamental theory, applications, algorithms, and implementations of processing or transferring information contained in many different physical, symbolic, or abstract formats broadly designated as signals.

It uses mathematical, statistical, computational, heuristic representations, formalisms, and techniques for representation, modeling, analysis, synthesis, discovery, recovery, sensing, acquisition, extraction, learning, security, or forensics.

Practice Questions

1. What are the various steps involved in Analysis of Data?
2. Provide Definition and examples for the following types of data.
    a. Nominal
    b. Ordinal
    c. Categorical
    d. Continues
    e. Discrete.
3. Explain the applications of Analytics in detail?
4. Discuss different data models in business domain applications?
5. What are the Applications of Modeling in Business?
6. Explain different data types used in data Analytics?
7. How to impute missing values by using Multiple Imputation by Chained Equation (MICE) Package ?
8. Explain about analytics applications to various business domains?
9. Define Predictive analytics? Discuss in detail?
10. Explain about Different Tools used for Analytics?
11. Explain about different types of data available in data analytics.
12. Discuss various Data analytics techniques with examples?