III B.Tech I Semester (R18) 2021-22

UNIT – II- Data Analytics

Data Analytics: Introduction to Analytics, Introduction to Tools and Environment, Application of Modelling in Business, Databases & Types of Data and variables, Data Modelling Techniques, Missing Imputations etc. Need for Business Modelling.

Data?

- In computing, data is information that has been translated into a form that is efficient for movement or processing.
- Data can exist in a variety of forms as numbers or text on pieces of paper, as bits and bytes stored in electronic memory, or as facts stored in a person's mind.
-

Analytics?

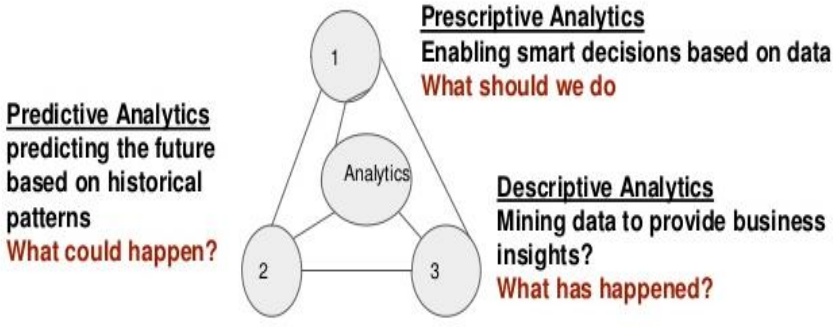
- Analytics is the discovery, interpretation, and communication of meaningful patterns in **data** and applying those patterns towards effective decision making.
- Analytics is an encompassing and multidimensional field that uses mathematics, statistics, predictive modelling and **machine learning** techniques to find meaningful patterns and knowledge in recorded data.

What is DATA Analytics?

- Data analysis is a process of inspecting, cleaning, transforming and modeling data.
- Data Analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain.

Different Stages Of Business Analytics

For different stages of business analytics huge amount of data is processed at various steps. Depending on the stage of the workflow and the requirement of data analysis, there are four main kinds of analytics – descriptive, diagnostic, predictive and prescriptive. These four types together answer everything a company needs to know- from what's going on in the company to what solutions to be adopted for optimising the functions.



The four types of analytics are usually implemented in stages and no one type of analytics is said to be better than the other. They are interrelated and each of these offers a different insight. With data being important to so many diverse sectors- from manufacturing to energy grids, most of the companies rely on one or all of these types of analytics. With the right choice of analytical techniques, big data can deliver richer insights for the companies

1) **Descriptive Analytics**: Describing or summarising the existing data using existing business intelligence tools to better understand what is going on or what has happened.

2) **Diagnostic Analytics**: Focus on past performance to determine what happened and why. The result of the analysis is often an analytic dashboard.

3) **Predictive Analytics**: Emphasizes on predicting the possible outcome using statistical models and machine learning techniques.

4) **Prescriptive Analytics**: It is a type of predictive analytics that is used to recommend one or more course of action on analyzing the data.

Let's understand these in a bit more depth.

Why Data Analytics

- Data Analytics is needed in Business to Consumer applications (B2C). Organizations collect data that they have gathered from customers, Business, economy and practical experience.
- Data is then processed after gathering and is categorized as per the requirement and analysis is done to study purchase patterns and etc.

Importance of Data Analytics

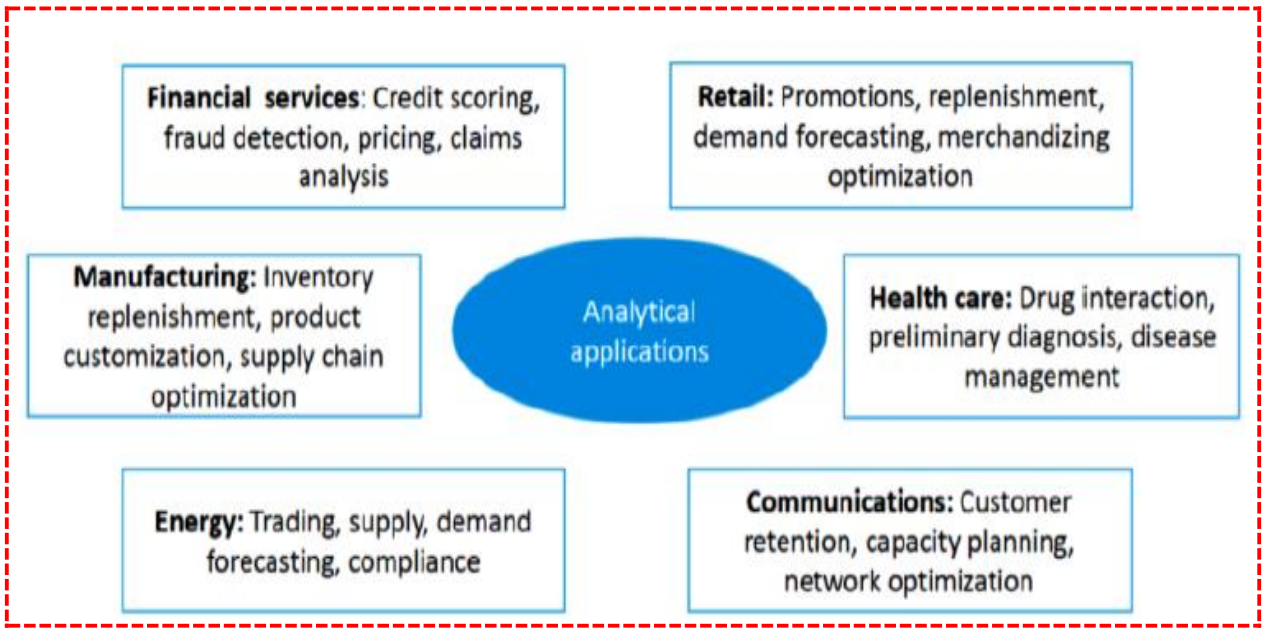
Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.

A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new—and better—products and services.

- ✓ Predict customer trends and behaviours
- ✓ Analyse, Interpret and deliver data in meaningful ways.
- ✓ Increase business productivity
- ✓ Drive effective decision-making

Places where Analytics is used

Some of the sectors that have adopted the use of data analytics include the travel and hospitality industry, where turnarounds can be quick. This industry can collect customer data and figure out where the problems, if any, lie and how to fix them.



Healthcare combines the use of high volumes of structured and unstructured data and uses data analytics to make quick decisions. Similarly, the retail industry uses copious amounts of data to meet the ever-changing demands of shoppers. The information retailers collect and analyze can help them identify trends, recommend products, and increase profits.

The Process of Data Analysis

- Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users.

- There are seven phases that can be distinguished:
  - ❖ Data requirements,
  - ❖ Data Collection,
  - ❖ Data Processing,
  - ❖ Data Cleaning,
  - ❖ Exploratory data analysis,
  - ❖ Modeling and Algorithms,
  - ❖ data product
  - ❖ Communication

Data Analytics Tools and Environment

The growing demand and importance of data analytics in the market have generated many openings worldwide. It becomes slightly tough to shortlist the top data analytics tools as the open source tools are more popular, user-friendly and performance oriented than the paid version. There are many open source tools which doesn't require much/any coding and manages to deliver better results than paid versions e.g. – R programming in data mining and Tableau public, Python in data visualization. Below is the list of top 10 of data analytics tools, both open source and paid version, based on their popularity, learning and performance.

With the increasing demand for Data Analytics in the market, many tools have emerged with various functionalities for this purpose. Either open source or User-Friendly, the top tools in the data analytics market are as follows.

R Programming: This tool is the leading analytics tool used for statistics and data modelling. R compiles and runs on various platforms such as UNIX, Windows, and Mac OS. It also provides tools to automatically install all packages as per user-requirement.

Python: Python is an open-Source, Object-oriented programming language which is easy to read, write and maintain. It provides various machine learning and visualization libraries such as Scikit-learn, TensorFlow, Matplotlib, Pandas, Keras etc. It also can be assembled on any platform like SQl server,a MongoDB database or JSON.

1. R Programming

R is the leading analytics tool in the industry and widely used for statistics and data modeling. It can easily manipulate your data and present in different ways. It has exceeded SAS in many ways like capacity of data, performance and outcome. R compiles and runs on a wide variety of platforms viz -UNIX, Windows and MacOS. It has 11,556 packages and allows you to browse the packages by categories. R also provides tools to automatically install all packages as per user requirement, which can also be well assembled with Big data.

2. Tableau Public:

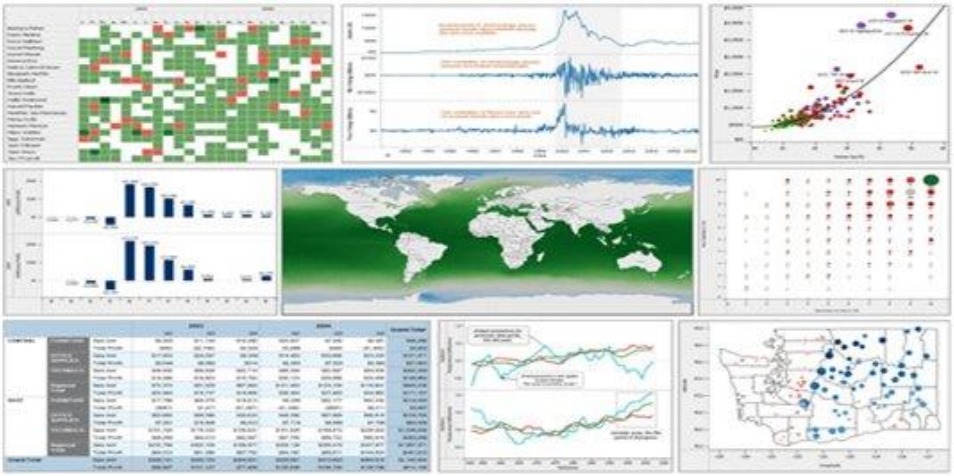
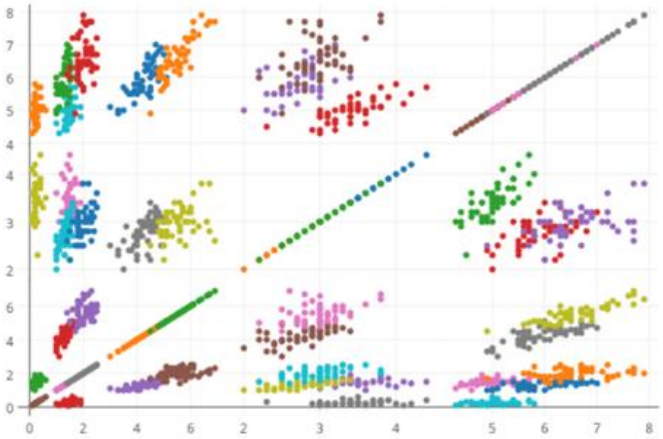


Tableau Public is a free software that connects any data source be it corporate Data Warehouse, Microsoft Excel or web-based data, and creates data visualizations, maps, dashboards etc. with real-time updates presenting on web. They can also be shared through social media or with the client. It allows the access to download the file in different formats. If you want to see the power of tableau, then we must have very good data source. Tableau's Big Data capabilities makes them important and one can analyze and visualize data better than any other data visualization software in the market.

3.Python

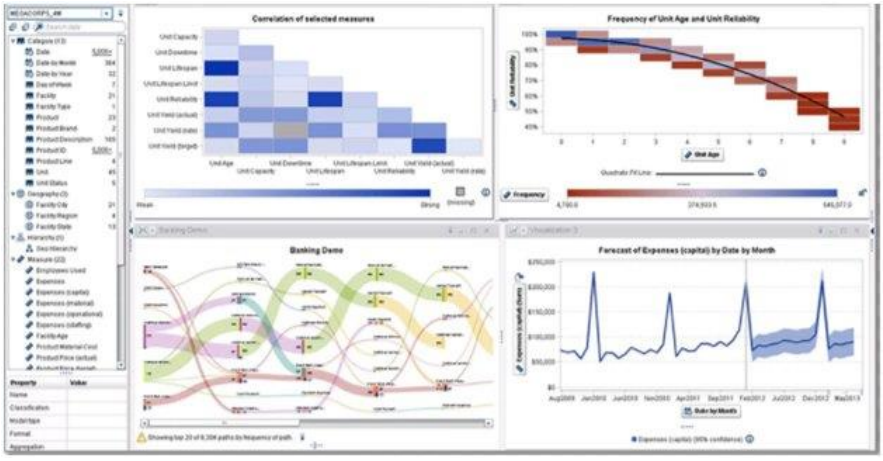
Python is an object-oriented scripting language which is easy to read, write, maintain and is a free open source tool. It was developed by Guido van Rossum in late 1980's which supports both functional and structured programming methods.



Phython is easy to learn as it is very similar to JavaScript, Ruby, and PHP. Also, Python has very good machine learning libraries viz. Scikitlearn, Theano, Tensorflow and Keras. Another important feature of Python is that it can be assembled on any platform like SQL server, a MongoDB database or JSON. Python can also handle text data very well.

4. SAS:

Sas is a programming environment and language for data manipulation and a leader in analytics, developed by the SAS Institute in 1966 and further developed in 1980's and 1990's. SAS is easily accessible, managable and can analyze data from any sources.



SAS introduced a large set of products in 2011 for customer intelligence and numerous SAS modules for web, social media and marketing analytics that is widely used for profiling customers and prospects. It can also predict their behaviors, manage, and optimize communications.

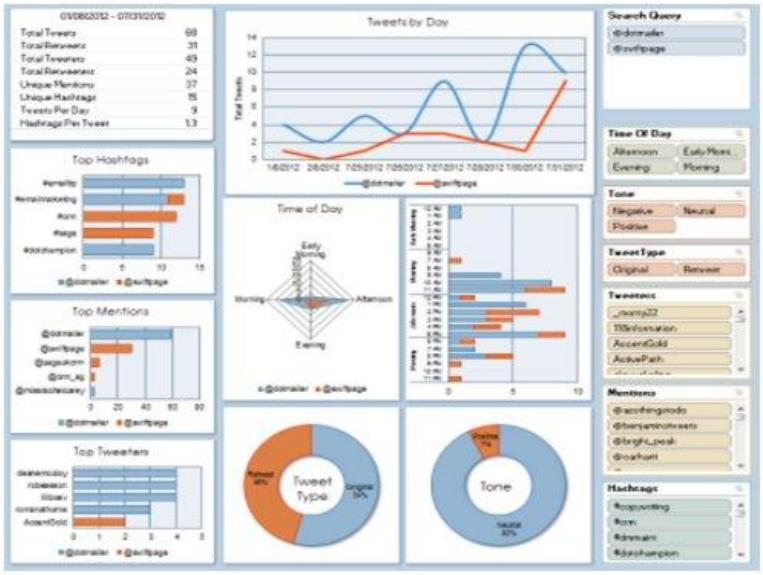
5. Apache Spark

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. Spark is built on data science and its concept makes data science effortless. Spark is also popular for data pipelines and machine learning models development.

Spark also includes a library – MLlib, that provides a progressive set of machine algorithms for repetitive data science techniques like Classification, Regression, Collaborative Filtering, Clustering, etc.

6. Excel

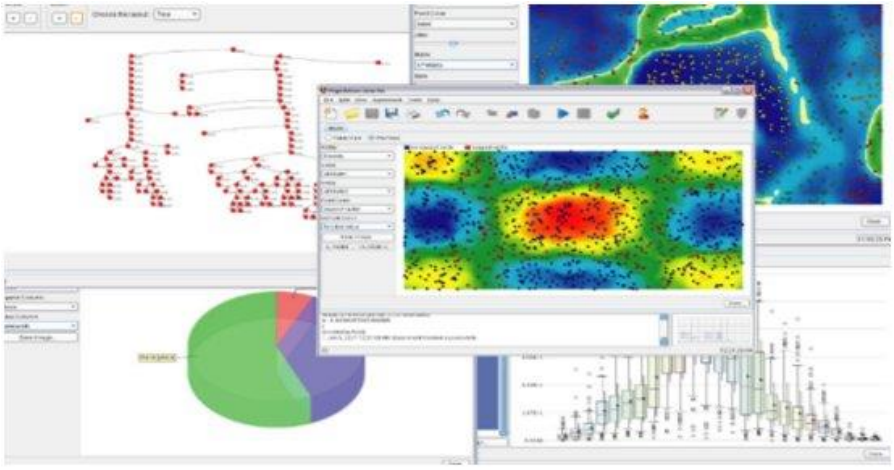
Excel is a basic, popular and widely used analytical tool almost in all industries. Whether you are an expert in Sas, R or Tableau, you will still need to use Excel. Excel becomes important when there is a requirement of analytics on the client's internal data. It analyzes the complex task that summarizes the data with a preview of pivot tables that helps in filtering the data as per client requirement.



Excel has the advance business analytics option which helps in modelling capabilities which have prebuilt options like automatic relationship detection, a creation of DAX measures and time grouping.

7. RapidMiner:

RapidMiner is a powerful integrated data science platform developed by the same company that performs predictive analysis and other advanced analytics like data mining, text analytics, machine learning and visual analytics without any programming. RapidMiner can incorporate with any data source types, including Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase, IBM DB2, Ingres, MySQL, IBM SPSS, Dbase etc.



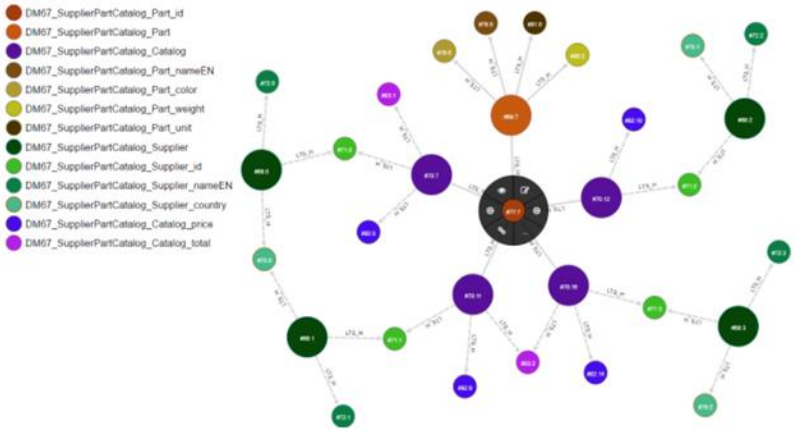
The tool is very powerful that can generate analytics based on real-life data transformation settings, i.e. you can control the formats and data sets for predictive analysis.

8. KNIME

KNIME Developed in January 2004 by a team of software engineers at University of Konstanz. KNIME is leading open source, reporting, and integrated analytics tools that allow you to analyze and model the data through visual programming, it integrates various components for data mining and machine learning via its modular data-pipelining concept.

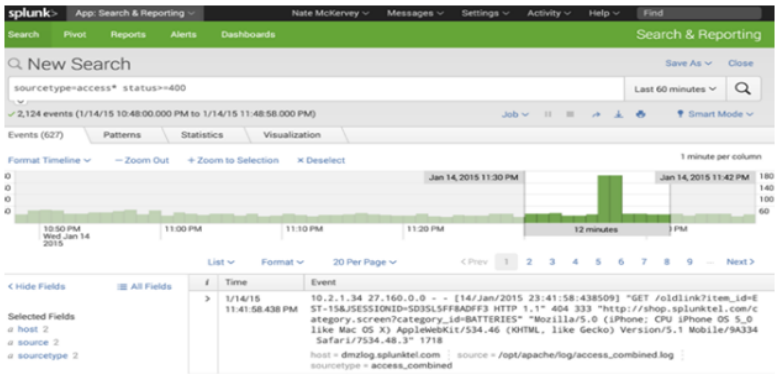
9. QlikView

QlikView has many unique features like patented technology and has in-memory data processing, which executes the result very fast to the end users and stores the data in the report itself. Data association in QlikView is automatically maintained and can be compressed to almost 10% from its original size. Data relationship is visualized using colors – a specific color is given to related data and another color for non-related data.



10. Splunk:

Splunk is a tool that analyzes and search the machine-generated data. Splunk pulls all text-based log data and provides a simple way to search through it, a user can pull in all kind of data, and perform all sort of interesting statistical analysis on it, and present it in different formats.



Top 10 Data Analytics Tools You Need To Know In 2020

The word 'Data' has been in existence for ages now. In the era of 2.5 Quintillion bytes of data being generated every day, data plays a crucial role in decision making for business operations. But how do you think we can deal with so much data? Well, there are several roles in the industry today that deal with data to gather insights, and one such vital role is of a Data Analyst. A Data Analyst requires many tools to gather insights from data. This article on the Top 10 Data Analytics Tools will talk about the top tools that every budding Data Analyst to a skilled professional must learn in 2020.

In this article, we will cover the following Data Analytics Tools:

1. R and Python
2. Microsoft Excel
3. Tableau
4. RapidMiner
5. KNIME
6. Power BI
7. Apache Spark
8. QlikView
9. Talend
10. Splunk

R and Python

R and Python are the top programming languages used in the Data Analytics field. R is an open-source tool used for Statistics and Analytics whereas Python is a high level, an interpreted language that has an easy syntax and dynamic semantics.

Products

Both R and Python are completely free and you can easily download both of them from their respective official websites.

Companies using

Companies such as ANZ, Google, Firefox use R, and other multinational companies such as YouTube, Netflix Facebook use Python.

Recent Advancements/ Features

Python and R are developing their features and functionalities to ease the process of Data Analysis with high speed and accuracy. They are coming up with various releases on a frequent basis with their updated features.

Microsoft Excel

Microsoft Excel is a platform that will help you get better insights into your data. Being one of the most popular tools for Data Analytics, Microsoft Excel provides the users with features such as sharing workbooks, work on the latest version for real-time collaboration, and adding data to Excel directly from a photo and so on.

Products

Microsoft Excel offers products in the following three categories:

- For Home
- For Business
- For Enterprises

Few of the versions are available for free for 1 month. All these products have various versions which differ by features and their pricing options.

Companies using

Almost all organizations use Microsoft Excel on a daily basis to gather meaningful insights from the data. A few of the popular names are McDonald's, IKEA, Marriot.

Recent Advancements/ Features

The recent advancements vary on the basis of the platform. Few of the recent advancements in Windows platform are as follows:

- You can get a snapshot of your workbook with Workbook Statistics
- You can give your docum

Tableau

Tableau is a market-leading Business Intelligence tool used to analyze and visualize data in an easy format. Being named as a **leader in the Gartner Magic Quadrant 2020 For the eighth consecutive year**, Tableau allows you to work on live data-set and spend more time on Data Analysis rather than Data Wrangling.

Products

Tableau Product Family include the following:

- Tableau Desktop
- Tableau Server
- Tableau Online
- Tableau Reader
- Tableau Public

Out of all, **Tableau Public is a free** Tableau software that you can use to make visualizations with but you need to save your workbook or worksheets in the Tableau Server which can be viewed by anyone.

Companies using

Multinational organizations such as Citibank, Deloitte, Skype, and Audi use Tableau to visualize their data and generate meaningful insights.

Recent Advancements/ Features

Tableau is coming up with frequent updates to provide users with the following:

- Fast Analytics
- Smart Dashboards
- Update Automatically
- Ease of Use
- Explore any data
- Publish a dashboard and share it live on the web and on mobile devices.

RapidMiner

RapidMiner is the next tool on our list. Being named a **Visionary in 2020 Gartner Magic Quadrant for Data Science and Machine Learning Platforms**, RapidMiner is a platform for data processing, building Machine Learning models, and deployment.

Products

The products of RapidMiner are as follows:

- Studio
- GO
- Server
- Real-Time Scoring
- Radoop

All these products have sub-versions which differ by features offered by them and pricing options.

Companies using

Companies such as BMW, Hewlett Packard Enterprise, EZCater, Sanofi use RapidMiner for their Data Processing and Machine Learning models.

Recent Advancements/ Features

Recently RapidMiner has launched **RapidMiner 9.6 which has extended the platform to full-time coders and BI Users**. It is a fully transparent, end-to-end Data Science platform that enables data preparation, Machine Learning, and model operations.

KNIME

Konstanz Information Miner or most commonly known as KNIME is free and an open-source data analytics, reporting, and integration platform built for analytics on a GUI based workflow.

Products

KNIME provides the following two software:

- **KNIME Analytics Platform** – Is an open-source and used to clean & gather data, make reusable components accessible to everyone, and create Data Science workflows.
- **KNIME Server** – Is a platform used by enterprises for the deployment of Data Science workflows, team collaboration, management, and automation.

Companies using

Companies such as Siemens, Novartis, Deutsche Telekom, Continental use KNime to make sense of their data and leverage meaningful insights.

Recent Advancements/ Features

You **do not need prior programming knowledge** to use KNIME and derive insights. You can work all the way from gathering data and creating models to deployment and production.

Power BI

Power BI is a Microsoft product used for business analytics. **Named as a leader for the 13th consecutive year in the Gartner 2020 Magic Quadrant,** it provides interactive visualizations with self-service business intelligence capabilities, where **end users can create dashboards and reports** by themselves, without having to depend on anybody.

Products

Power BI provides the following products:

- Power BI Desktop
- Power BI Pro
- Power BI Premium
- Power BI Mobile
- Power BI Embedded
- Power BI Report Server

All these products differ by the functionalities offered by them. Few of them are free for a certain period of time and then you have to take the licensed versions

Companies using

Multinational organizations such as Adobe, Heathrow, Worldsmart, GE Healthcare are using Power BI to achieve powerful results from their data.

Recent Advancements/ Features

Power BI has recently come up with solutions such as **Azure + Power BI** and **Office 365 + Power BI** to help the users analyze the data, connect the data and protect the data across various Office platforms.

Apache Spark

Apache Spark is one of the most successful projects in the Apache Software Foundation and is a *cluster computing framework* that is **open-source and is used for real-time processing.** Being the most active Apache project at the moment, it comes with a fantastic **open-source community** and an **interface for programming**. This interface makes sure of fault tolerance and implicit data parallelism.

Products

Apache Spark keeps on releasing new releases with new features. You can also choose the various package types for Spark. The **recent version is 2.4.5** and **3.0.0 is in preview.**

Companies using

Companies such as Oracle, Hortonworks, Verizon, Visa use Apache Spark for real-time computation of data with ease of use and speed.

Recent Advancements/ Features

- In today's world Spark runs on Kubernetes, Apache Mesos, standalone, Hadoop, or in the cloud.

- It provides high-level APIs in Java, Scala, Python, and R, and Spark code can be written in any of these four languages.
- Spark's MLlib – the Machine Learning component is handy when it comes to Big Data processing.

QlikView

QlikView is a **Self-Service Business Intelligence, Data Visualization, and Data Analytics tool**. Being named a **leader in Gartner Magic Quadrant 2020 for Analytics and BI platforms,** it aims to accelerate business value through data by providing features such as Data Integration, Data Literacy, and Data Analytics.

Products

QlikView comes with a variety of products and services for **Data Integration, Data Analytics**, and **Developer platforms**, out of which few are available for a **free trial period of 30 days**.

Companies using

Trusted by more than 50,000 customers worldwide few of the top customers of QlikView are CISCO, NHS, KitchenAid, SAMSUNG.

Recent Advancements/ Features

Recently QlikView has launched an **intelligent alerting platform Qlik Alerting for Qlik Sense®** which helps the organizations handle the exceptions, notify users of potential issues, help users analyze further, and also prompts actions based on the derived insights.

Talend

Talend is one of the most powerful data integration ETL tools available in the market and is developed in the Eclipse graphical development environment. Being named as a **Leader in Gartner's Magic Quadrant for Data Integration Tools and Data Quality tools 2019**, this tool lets you **easily manage all the steps involved in the ETL process** and aims to deliver compliant, accessible and clean data for everyone.

Products

Talend comes with the following five products:

- Talend Open Source
- Stitch Data Loader
- Talend Pipeline Designer
- Talend Cloud Data Integration
- Talend Data Fabric

Out of these, few are completely free, few are free for 14 days and few are licensed. All these products differ in their functionalities and pricing options.

Companies using

Small startups to multinational companies such as ALDO, ABInBev, EuroNext, AstraZeneca are using Talend to make critical decisions.

Recent Advancements/ Features

Talend is the only platform that **delivers complete and clean data at the moment you need it** by maintaining data quality, providing Big Data integration, cloud API services, Preparing Data, and providing Data Catalog and Stitch Data Loader.

Splunk

Splunk is a platform used to **search, analyze**, and **visualize the machine-generated data** gathered from the applications, websites, etc. Being named by **Gartner as a Visionary in the 2020 Magic Quadrant for APM**, Splunk has evolved products in various fields such as IT, Security, DevOps, Analytics.

Products

- Splunk Free
- Splunk Enterprise
- Splunk Cloud

All these 3 products differ by the bandwidth of the features they offer and are available for free download and trial versions. The pricing options for Splunk products are based on predictive pricing, Infrastructure-based pricing, and also rapid adoption packages.

Companies using

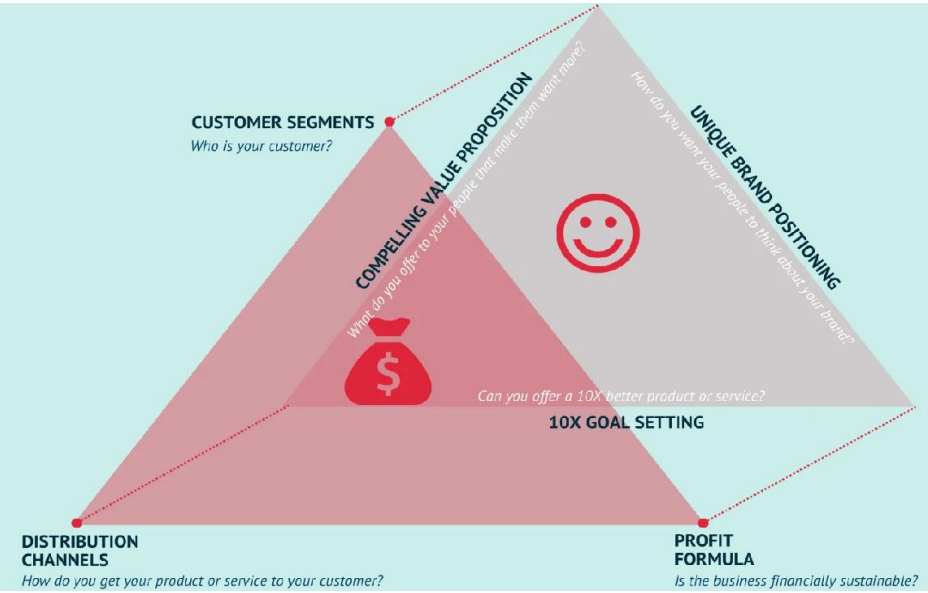
Trusted by 92 out of the Fortune 100**, companies such as Dominos, Otto Group, Intel, Lenovo are using Splunk in their day to day practices to discover the processes and correlate data in real-time.**

Recent Advancements/ Features

Since almost all the organizations need to deal with data across various divisions, according to Splunk official website Splunk aims to bring data to every part of your organization, by helping teams use Splunk to **prevent and predict problems with monitoring experience, detect and diagnose issues with clear visibility, explore** and **visualize business processes and streamline the entire security stack**.

Business Model

- A business model is a conceptual structure that supports the viability of the business and explains how it operates, makes money, and how it intends to achieve its goals.
- All the business processes and policies that a company adopts and follows are part of the business model.



- ***"A business model is supposed to answer who your customer is, what value you can create/add for the customer and how you can do that at reasonable costs".***
- **Thus, a business model is a description of how a company creates, delivers, and captures value for itself as well as the customer.**

In more simple terms, every business model intrinsically has three parts –

- **Everything related to designing and manufacturing the product.**
- **Everything related to selling the product, from finding the right customers to distributing the product.**
- **Everything related to how the customer will pay and how the company will make money.**

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.

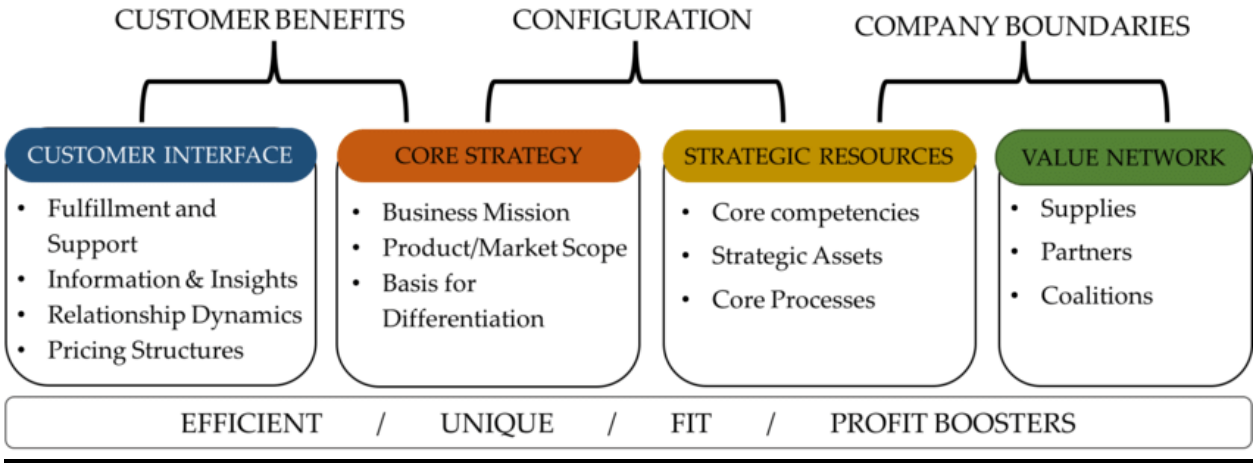
- ✓ Signal processing is an enabling technology that encompasses the fundamental theory, applications, algorithms, and implementations of processing or transferring information contained in many different physical, symbolic, or abstract formats broadly designated as signals.

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

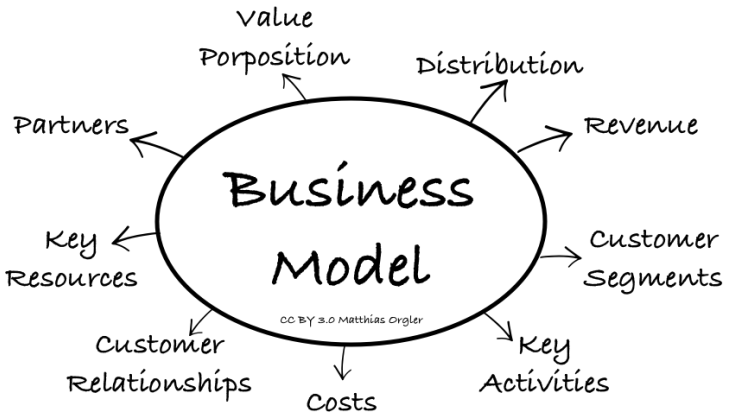
- ✓ In manufacturing statistical models are used to define Warranty policies, solving various conveyor related issues, Statistical Process Control etc.

Important Components of Business Model

- The business model acts as the blueprint of the business and a roadmap for its success (or failure) as it explains how the business creates and capture value through its decisions and processes.



- Creating a business model isn't simply about completing your business plan or determining which products to pursue. It's about mapping out how you will create ongoing value for your customers.



1. Identify your specific audience.

Targeting a wide audience won't allow your business to hone in on customers who truly need and want your product or service. Instead, when creating your business model, narrow your audience down to two or three detailed buyer personas. Outline each persona's demographics, common challenges and the solutions your company will offer.

2. Establish business processes.

Before your business can go live, you need to have an understanding of the activities required to make your business model work. Determine key business activities by first identifying the core aspect of your business's offering.

3. Record key business resources.

What does your company need to carry out daily processes, find new customers and reach business goals? Document essential business resources to ensure your business model is adequately prepared to sustain the needs of your business.

4. Develop a strong value proposition.

How will your company stand out among the competition? Do you provide an innovative service, revolutionary product or a new twist on an old favorite? Establishing exactly what your business offers and why it's better than competitors is the beginning of a strong value proposition.

5. Determine key business partners.

No business can function properly (let alone reach established goals) without key partners that contribute to the business's ability to serve customers. When creating a business model, select key partners, like suppliers, strategic alliances or advertising partners.

6. Create a demand generation strategy.

Unless you're taking a radical approach to launching your company, you'll need a strategy that builds interest in your business, generates leads and is designed to close sales.
Databases & Type of data and variables

7. Leave room for innovation.

When launching a company and developing a business model, your business plan is based on many assumptions. After all, until you begin to welcome paying customers, you don't truly know if your business model will meet their ongoing needs. For this reason, it's important to leave room for future innovations. Don't make a critical mistake by thinking your initial plan is a static document. Instead, review it often and implement changes as needed.

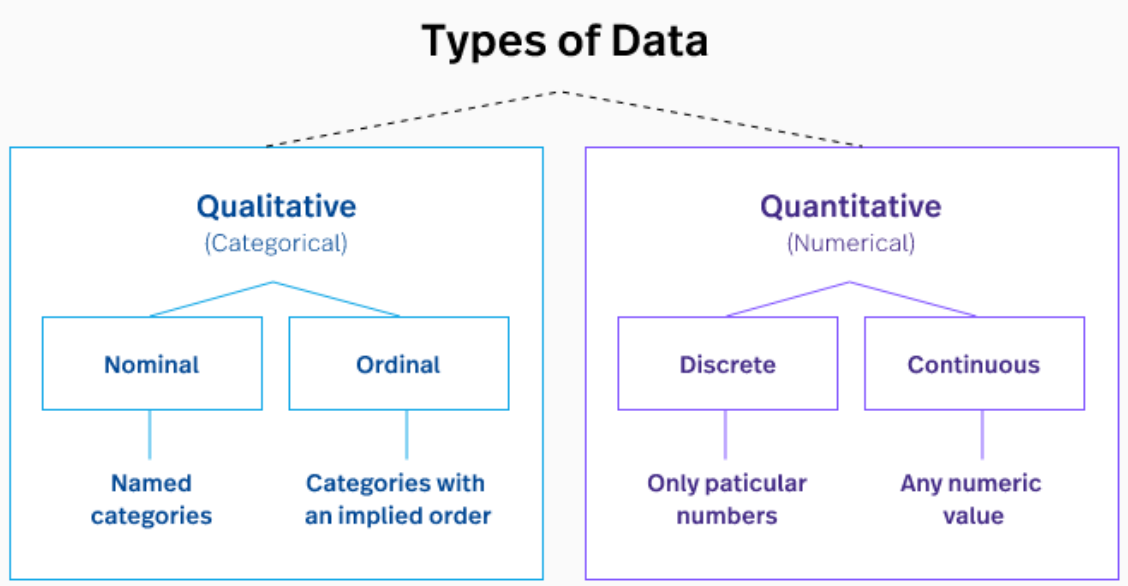
Data Dictionary:

Data Dictionary is a central source of information for the data in a information management system. Its main function is to support the creation and management of data definitions (or "metadata").

Data dictionary refers to a specific place in a database that stores information related to different types of data present in all the databases. It can determine relevant structure of data and restructures the data query. the information stored in a data dictionary consists of tables, indexes, schemas, views, functions, procedures, user permissions, user statistics, database design, database growth, performance and so on.

- It refers to a specific place in a database that stores information related to different types of data present in all the databases.
- It can identify relevant structure of data and restructure the data query.

Data is all around us, and every day it becomes increasingly important. Different types of data define more and more of our interactions with the world around us—from using the internet, to buying a car, to the algorithms behind news feeds we see, and much more.



Quantitative Variables - Variables whose values result from counting or measuring something.

Examples:

- Number of students in a class
- Number of square feet in a house
- Population size of a city
- Age of an individual
- Height of an individual

Qualitative Variables - Variables that are not measurement variables. Their values do not result from measuring or counting.

Examples:

- Eye color (e.g. “blue”, “green”, “brown”)
- Gender (e.g. “male”, “female”)
- Breed of dog (e.g. “lab”, “bulldog”, “poodle”)
- Level of education (e.g. “high school”, “Associate’s degree”, “Bachelor’s degree”)
- Marital status (e.g. “married”, “single”, “divorced”)

Every single variable you will ever encounter in statistics can be classified as either quantitative or qualitative.

	Quantitative Variables	Qualitative Variables
Definition	*Take on numeric values*	*Take on names or labels*
Examples	Number of students in a class	Eye color
	Number of square feet in a house	Gender
	Population size of a city	Breed of dog
	Age of an individual	Level of Education
	Height of an individual	Marital status

There are five total variables in this dataset. Two of them are qualitative variables and three of them are quantitative variables:

Variable Type:	Qualitative	Qualitative	Quantitative	Quantitative	Quantitative
Player Name	**Position**	**Seasons Played**	**Avg. Points**	**Championships**
Mike	G	12	22.1	3
Chuck	G	9	26.6	2
Tony	F	8	16.5	2
Andy	F	8	17.7	0
Karl	C	14	24.4	1
John	G	12	29.8	2
Klay	F	16	17.2	2
Dirk	F	15	14.4	4
Mark	G	9	9.8	3
Kenny	C	12	20.1	3

Qualitative

Qualitative variables are divided into two types: **nominal** and **ordinal**.

Nominal

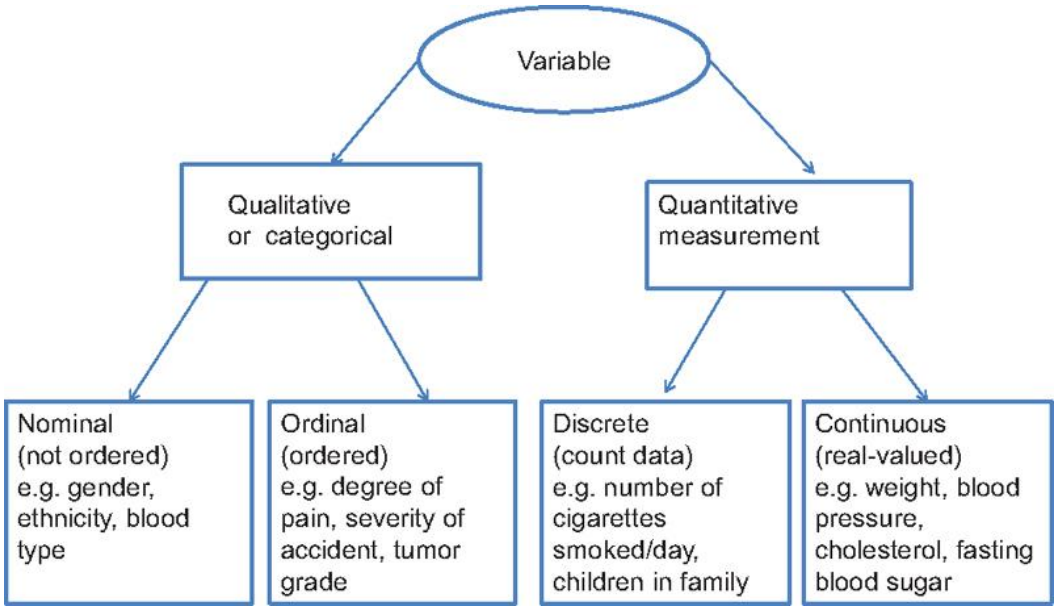
A **qualitative nominal** variable is a qualitative variable where **no ordering** is possible or implied in the levels. For example, the variable gender is nominal because there is no order in the levels female/male. Eye color is another example of a nominal variable because there is no order among blue, brown or green eyes.

A nominal variable can have between two levels (e.g., do you smoke? Yes/No or what is your gender? Female/Male) and a large number of levels (what is your college major? Each major is a level in that case).

Ordinal

On the other hand, a **qualitative ordinal** variable is a qualitative variable with an **order implied in the levels**. For instance, if the severity of road accidents has been measured on a scale such as light, moderate and fatal accidents, this variable is a qualitative ordinal variable because there is a clear order in the levels.

Another good example is health, which can take values such as poor, reasonable, good, or excellent. Again, there is clear order in these levels so health is in this case a qualitative ordinal variable.



Quantitative

Quantitative variables are divided into two types: **discrete** and **continuous**. The difference is explained in the following two sections.

Discrete

Quantitative discrete variables are variables for which the values it can take are **countable** and have a **finite number of possibilities**. The values are often (but not always) integers. Here are some examples of discrete variables:

- Number of children per family
- Number of students in a class
- Number of citizens of a country

Even if it would take a long time to count the citizens of a large country, it is still technically doable. Moreover, for all examples, the number of possibilities is **finite**. Whatever the number of children in a family, it will never be 3.58 or 7.912 so the number of possibilities is a finite number and thus countable.

Continuous

On the other hand, **quantitative continuous** variables are variables for which the values are **not countable** and have an **infinite number of possibilities**. For example:

- Age
- Weight
- Height

For simplicity, we usually referred to years, kilograms (or pounds) and centimeters (or feet and inches) for age, weight and height respectively. However, a 28-year-old man could actually be 28 years, 7 months, 16 days, 3 hours, 4 minutes, 5 seconds, 31 milliseconds, 9 nanoseconds old.

For all measurements, we usually stop at a standard level of granularity, but nothing (except our measurement tools) prevents us from going deeper, leading to an **infinite number of potential values**. The fact that the values can take an infinite number of possibilities makes it uncountable.

Data Modeling Techniques

- ✓ Regression analysis mainly focuses on finding a relationship between a dependent variable and one or more independent variables.
- ✓ Predict the value of a dependent variable based on the value of at least one independent variable.
- ✓ It explains the impact of changes in an independent variable on the dependent variable.
- ✓ $Y = f(X, \beta)$ where Y is the dependent variable X is the independent variable β is the unknown coefficient.
- ✓ Widely used in prediction and forecasting.

- **Data modeling** defines not just **data** elements, but also their structures and the relationships between them. **Data modeling techniques** and methodologies are used to **model data** in a standard, consistent, predictable manner in order to manage it as a resource.

The following are the various data modeling techniques

- ❖ Regression
- ❖ Classification
- ❖ Clustering
- ❖ Anomaly Detection

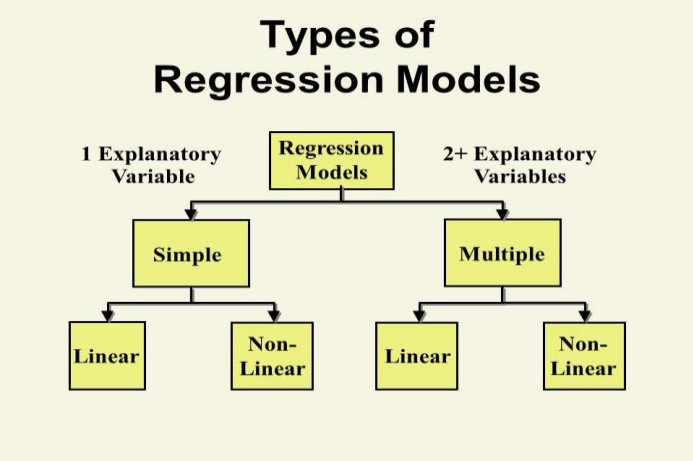
Regression

Regression is a method to mathematically formulate relationship between variables that in due course can be used to estimate, interpolate and extrapolate. Suppose we want to estimate the weight of individuals, which is influenced by height, diet, workout, etc. Here, Weight is the predicted variable. *Height, Diet, Workout* is predictor variables.

The predicted variable is a **dependant** variable in the sense that it depends on predictors. Predictors are also called as **independent** variables. Regression reveals to what extent the predicted variable is affected by the predictors. In other words, what amount of variation in predictors will result in variations of the predicted

variable. The predicted variable is mathematically represented as YY. The predictor variables are represented as X1, X2, X3, etc. This mathematical relationship is often called the **regression model**.

Regression is a branch of statistics. There are many types of regression. Regression is commonly used for prediction and forecasting.



When it comes to the level of analysis in statistics, there are three different analysis techniques that exist. These are –

- Univariate analysis
- Bivariate analysis
- Multivariate analysis

The selection of the data analysis technique is dependent on the number of variables, types of data and focus of the statistical inquiry.
The following section describes the three different levels of data analysis –

Univariate analysis

Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis technique is used.

Here is one example of Univariate analysis-

In a survey of a class room, the researcher may be looking to count the number of boys and girls. In this instance, the data would simply reflect the number, i.e. a single variable and its quantity as per the below table. The key objective of Univariate analysis is to simply describe the data to find patterns within the data. This is be done by looking into the mean, median, mode, dispersion, variance, range, standard deviation etc.

Univariate analysis is conducted through several ways which are mostly descriptive in nature –

- Frequency Distribution Tables
- Histograms

- Frequency Polygons
- Pie Charts
- Bar Charts

Bivariate analysis

Bivariate analysis is slightly more analytical than Univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data set then Bivariate analysis is the right type of analysis technique.

Here is one simple example of bivariate analysis –

In a survey of a classroom, the researcher may be looking to analysis the ratio of students who scored above 85% corresponding to their genders. In this case, there are two variables – gender = X (independent variable) and result = Y (dependent variable). A Bivariate analysis is will measure the correlations between the two variables.

Bivariate analysis is conducted using –

- Correlation coefficients
- Regression analysis

Multivariate analysis

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set.

Here is an example of multivariate analysis –

A doctor has collected data on cholesterol, blood pressure, and weight. She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week). She wants to investigate the relationship between the three measures of health and eating habits?

In this instance, a multivariate analysis would be required to understand the relationship of each variable with each other.
Commonly used multivariate analysis technique include –

- Factor Analysis
- Cluster Analysis
- Variance Analysis
- Discriminant Analysis
- Multidimensional Scaling
- Principal Component Analysis
- Redundancy Analysis

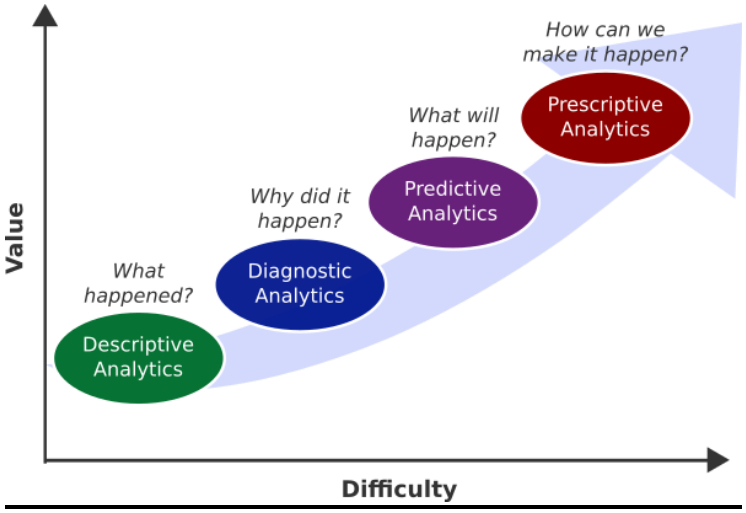
Classification

- Classification can be performed on structured or unstructured data. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under.

Applications of Classification Model

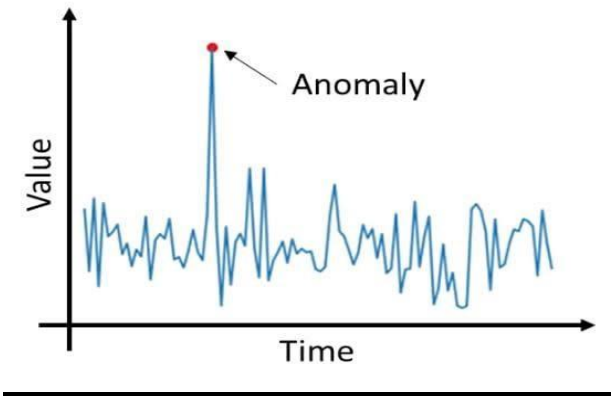
The classification model can be used in

- Modeling of Business
- Segmenting of customer
- Analyzing of credit



Anomaly Detection

- Anomaly detection (aka outlier analysis) is a step in data mining that identifies data points, events, and/or observations that deviate from a dataset's normal behavior. Anomalous data can indicate critical incidents, such as a technical glitch, or potential opportunities, for instance a change in consumer behavior.



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). Unlike SAS, R uses the same symbol for character and numeric data.
- To test if there is any missing in the dataset we use is.na () function.
- For Example, We have defined "y" and then checked if there is any missing value.
- T or True means that there is a missing value. y <- c(1,2,3,NA) is.na(y) # returns a vector (F FF T)
- Arithmetic functions on missing values yield missing values.
- For Example, x <- c(1,2,NA,3) mean(x) # returns NA 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)
- 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
- MICE Package -> Multiple Imputation by Chained Equations MICE uses PMM to impute missing values in a dataset.
- PMM-> Predictive Mean Matching (PMM) is a semi-parametric imputation approach.
- It is similar to the regression method except that for each missing value, it fills in a value randomly from among the observed donor values from an observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model.

In statistics, **imputation** is the process of replacing **missing** data with substituted values. That is to say, when one or more values are **missing** for a case, most statistical packages default to discarding any case that has a **missing** value, which may introduce bias or affect the representativeness of the results.

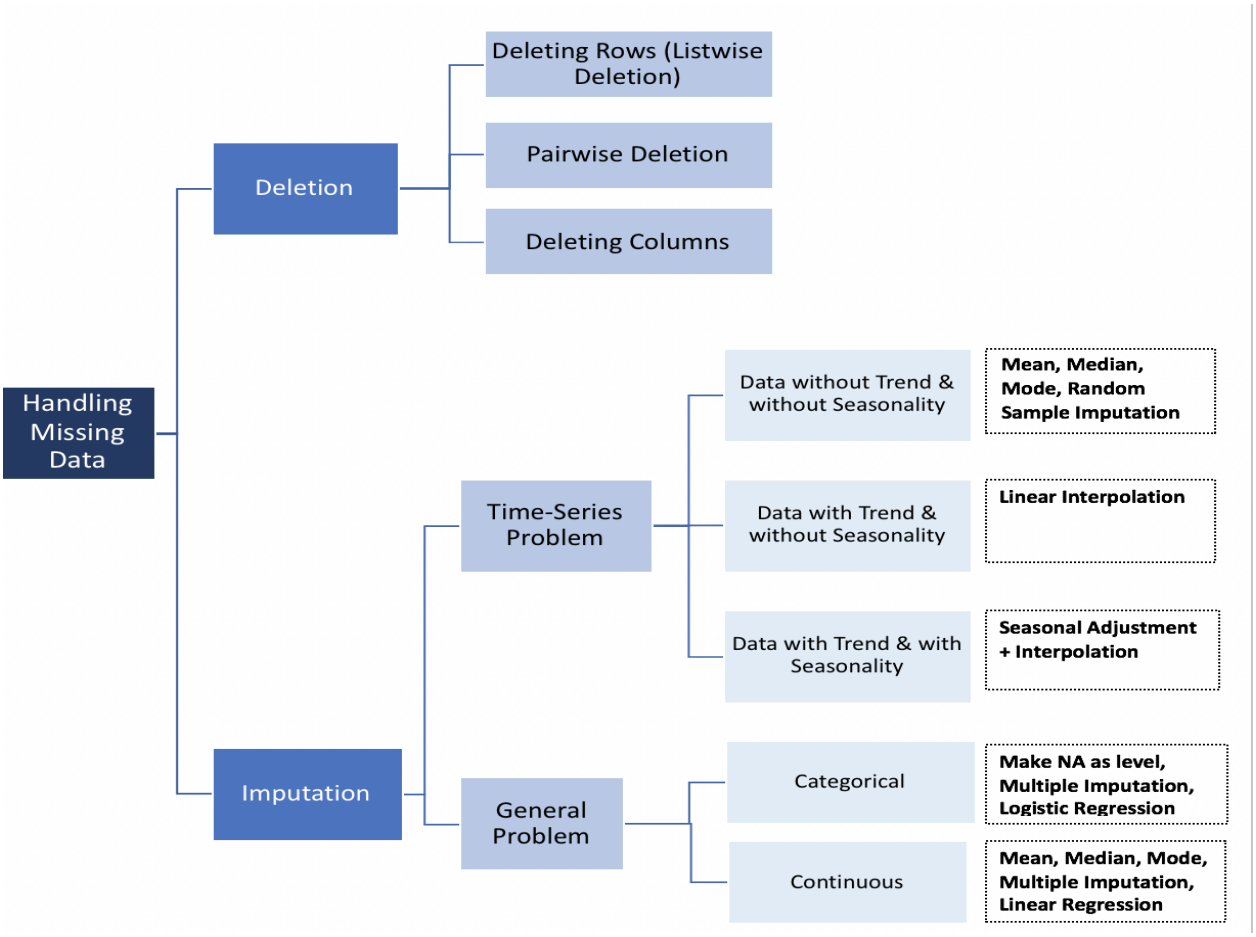
	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

- Many real-world datasets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders. Training a model with a dataset that has a lot of missing

values can drastically impact the machine learning model's quality. Some algorithms such as *scikit-learn estimators* assume that all values are numerical and have and hold meaningful value.

- **Missing completely at random (MCAR)**
- **Missing at random (MAR)**
- **Not missing at random (NMAR)**



Need for Business Modeling

Every business or companies makes a plan for generating profit. They create a model for identifying products and services to sell, the market they want to target and also take into account anticipated expenses. This is known as business models.

Even if the business is already established or even if it is a new business, plan needs to be made. Businesses need to regularly update their plans and strategy as they need to take into accounts the challenges and trends for the future models.

- Business Modeling can be defined as the process of exploring range of business decisions and identifying the elements that are essential for driving a business.

- Business Modeling assists from the very initial stage of idea generation to the final stage in the following manner.

✓ **Define**
✓ **Discover**
✓ **Develop**
✓ **Deliver**

Advantages of Business Models

- A good business models gives the company a competitive edge in the industry.
- A strong business model provides the company good reputation in the market place encouraging the investors to remain invested in the company.
- Making the business model strong leads to an ongoing business profit leading to increase in cash reserve and new investments.
- Proven business model brings a financial stability in the organization.