

Introduction to Data Science



Introduction to Data Science

Programs Offered

Post Graduate Programmes (PG)

- Master of Business Administration
- Master of Computer Applications
- Master of Commerce (Financial Management / Financial Technology)
- Master of Arts (Journalism and Mass Communication)
- Master of Arts (Economics)
- Master of Arts (Public Policy and Governance)
- Master of Social Work
- Master of Arts (English)
- Master of Science (Information Technology) (ODL)
- Master of Science (Environmental Science) (ODL)

Diploma Programmes

- Post Graduate Diploma (Management)
- Post Graduate Diploma (Logistics)
- Post Graduate Diploma (Machine Learning and Artificial Intelligence)
- Post Graduate Diploma (Data Science)

Undergraduate Programmes (UG)

- Bachelor of Business Administration
- Bachelor of Computer Applications
- Bachelor of Commerce
- Bachelor of Arts (Journalism and Mass Communication)
- Bachelor of Arts (General / Political Science / Economics / English / Sociology)
- Bachelor of Social Work
- Bachelor of Science (Information Technology) (ODL)



AMITY UNIVERSITY

DIRECTORATE OF
DISTANCE & ONLINE EDUCATION

Amity Helpline: 1800-102-3434 (toll-free), 0120-4614200

For Distance Learning Programmes: dladmissions@amity.edu | www.amity.edu/addoe

For Online Learning programmes: elearning@amity.edu | www.amityonline.com



AMITY

Introduction to Data Science



AMITY | DIRECTORATE OF DISTANCE &
UNIVERSITY ONLINE EDUCATION

© Amity University Press

All Rights Reserved

No parts of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior permission of the publisher.

SLM & Learning Resources Committee

Chairman : Prof. Abhinash Kumar

Members : Dr. Divya Bansal
Dr. Coral J Barboza
Dr. Monica Rose
Dr. Winnie Sharma

Member Secretary : Ms. Rita Naskar

Contents

Module - I: Introduction to Data Science

Page No.

01

- 1.1 Introduction to Data Science
 - 1.1.1 Definition of Data Science
 - 1.1.2 Benefits and Uses of Data Science
 - 1.1.3 Role of Data Scientist
- 1.2 Big Data and Data Science
 - 1.2.1 Definition: What is Big Data
 - 1.2.2 Evolution of Big Data and its Importance
 - 1.2.3 Four V's in Big Data, Drivers of Big Data
 - 1.2.4 Big Data and Data Science Hype; Datafication
- 1.3 Statistical Inferences
 - 1.3.1 Role of Statistics in Data Science, Inferences Types
 - 1.3.3 Population and Samples, Statistical Modelling
 - 1.3.4 Probability Distribution: Types and Role
 - 1.3.5 Fitting a Model
- 1.4 Introduction to R and Information Visualisation
 - 1.4.1 R Windows Environment, its Data Type, Functions, Loops, Data Structure
 - 1.4.2 R -Packages, Dataset Reading, Programming, Statistical Introduction
 - 1.4.3 Importance of Data Visualisation, Visualisation Aesthetics
 - 1.4.4 Proper Scaling and Colour, Effective Colour and Shading

Module - II: Exploratory Data Analysis and Data Science Process

105

- 2.1 Philosophy of Exploratory Data Analysis - The Data Science Process
 - 2.1.1 Descriptive Statistics and Data Preparation
 - 2.1.2 Exploratory Data Analysis-Summarisation, Measuring Asymmetry
 - 2.1.3 Sample and Estimated Mean, Variance, Standard Score
 - 2.1.4 Statistical inference, Frequency Approach, Variability of Estimates
 - 2.1.5 Hypothesis Testing
- 2.2 Basic Tools of EDA (Plots, Graphs and Summary Statistics)
 - 2.2.1 Chart Types: Tabular Data, Dot and Line Plot, Scatter plots, Bar plots, Pie Charts, Graphs
 - 2.2.2 Description of Data Using These Tools With Real time Example
- 2.3 Basic Data Science Process
 - 2.3.1 Overview of Data Science Process: Defining its Goal
 - 2.3.2 Retrieving the Data, Data Preparation-Exploration, Cleaning and Transforming Data
 - 2.3.3 Building the Model

- 2.3.4 Presentation and Automation
- 2.4 Machine Learning
 - 2.4.1 Introduction and Types of Machine Learning
 - 2.4.2 Role of Machine Learning in Data Science
 - 2.4.3 Classification Algorithms:-Linear Regression, Decision Tree
 - 2.4.4 Naive Bayes Classifier, K-means
 - 2.4.5 K-Nearest Neighbour, Support Vector Machine

Module - III: Feature Selection Algorithms

186

- 3.1 Feature Generation
 - 3.1.1 Extracting Feature from Data
 - 3.1.2 Transforming Features
 - 3.1.3 Selecting Features
 - 3.1.4 Role of Domain Expertise
- 3.2 Feature Selection Algorithms
 - 3.2.1 What is Feature Selection?
 - 3.2.2 Different Types of Feature Selection Methods
 - 3.2.3 Filter Methods: Types and Role
 - 3.2.4 Wrapper Method: Its Different Types
 - 3.2.5 Decision Tree: Its Importance and Role in Data Science
 - 3.2.6 Random Forest: Its Significance

Module - IV: Recommendation Systems

209

- 4.1 Dimensionality Reduction
 - 4.1.1 What is Predictive Modelling?
 - 4.1.2 What is Dimensionality Reduction?
 - 4.1.3 Importance of Dimensionality Reduction
 - 4.1.4 Different Components of Dimensionality Reduction
- 4.2 Singular Value Reduction
 - 4.2.1 Need for Dimensionality Reduction
 - 4.2.2 Understanding Singular Value Reduction: Mathematical Concept
 - 4.2.3 Single Value Theorem
- 4.3 Principal Component Analysis
 - 4.3.1 What is PCA?
 - 4.3.2 Algorithm for PCA
 - 4.3.3 Application and Role of PCA in Dimensionality Reduction
 - 4.3.4 Example for Finding PCA in Dataset

- 5.1 Text Mining and Information Retrieval
 - 5.1.1 Introduction to Text Mining
 - 5.1.2 Definition and Language for Data Science
 - 5.1.3 Collection of Data-Hunting, Logging, Scrapping
 - 5.1.4 Cleaning Data-Artifacts, Data Compatibility
 - 5.1.5 Dealing with Missing Values, Outliers
- 5.2 Big Data Fundamentals and Hadoop Integration with R
 - 5.2.1 Definition, Evolution of Big Data and its Importance
 - 5.2.2 Four Vs in Big Data, Drivers for Big data
 - 5.2.3 Big Data Analytics, Big Data Applications
 - 5.2.4 Designing Data Architecture, R Syntax
 - 5.2.5 IDE for Hadoop, Integration with Big Data, Integration Methods
- 5.3 Introduction to Neural Networks
 - 5.3.1 Introduction to Neural Network
 - 5.3.2 Difference Between Human Brain and Artificial Network
 - 5.3.3 Perceptron Model: Its Features, McCulloch-Pitts Model
 - 5.3.4 Role of Activation Function, Backpropagation Algorithm
 - 5.3.5 Neural Network in Data Science
- 5.4 Data Science and Ethical Issues
 - 5.4.1 Role of FAT in Data Science, Ethical Challenges in Data Science
 - 5.4.2 Some Real life Examples :Covid19, Data breach Cases

(c) Amity University Online

Module - I: Digital Marketing Fundamentals

Notes

Learning Objectives

At the end of this topic, you will be able to understand:

- Analyse definition of data science
- Describe benefits and uses of data science
- Identify role of data scientist
- Analyse big data and data science
- Describe evolution of big data and its importance
- Interpret four V's in big data, drivers of big data
- Analyse big data and data science hype; datafication
- Describe statistical inferences and role of statistics in data science, inferences types
- Analyse population and samples, statistical modelling
- Describe probability distribution: types and role
- Analyse fitting a model
- Identify introduction to R and information visualisation
- Describe R windows environment, its data type, functions, loops, data structure
- Analyse R -packages, dataset reading, programming, statistical introduction
- Interpret importance of data visualisation, visualisation aesthetics
- Identify proper scaling and colour, effective colour and shading

Introduction

The study of data in order to derive useful insights for businesses is referred to as “data science.” To analyse vast volumes of data, this method takes a multidisciplinary approach by combining concepts and methods from the domains of mathematics, statistics, artificial intelligence and computer engineering. The results of this analysis allow data scientists to ask and answer questions such as what occurred, why it occurred, what will occur and what can be done with the information.

1.1 Introduction to Data Science

In order to “understand and analyse actual events” with the use of data, a “concept to unite statistics, data analysis and informatics and their related approaches” has been developed called “data science.” Under the framework of mathematics, statistics, computer science, information science and domain knowledge, it makes use of techniques and theories borrowed from a wide variety of subjects. The study of data, on the other hand, is distinct from computer science and information science. The recipient of the Turing Award, Jim Gray, asserted that “everything about science is changing because of the impact of information technology” and the data deluge. Gray envisioned

Notes

data science as a “fourth paradigm” of science, which would follow empirical, theoretical, computational and now data-driven approaches to scientific enquiry.

1.1.1 Definition of Data Science

Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data. This can be accomplished by using statistical modelling, scientific computing, scientific methods, processes, algorithms and systems. In addition to this, domain knowledge from the underlying application domain is included into data science (e.g., natural sciences, information technology and medicine). Data science is a multidimensional field that encompasses not only a science but also a research paradigm, a research technique, an academic discipline, a process and a professional occupation.

The Data Science Lifecycle

Now that you have an understanding of what data science is, we will go on to discussing the data science lifecycle. The lifespan of data science is comprised of five separate stages, each of which is responsible for a unique set of tasks:

1. **Capture:** It consists of the following processes: data acquisition, data entry, signal reception and data extraction. During this stage, you will be acquiring raw data, both organised and unstructured.
2. **Maintain:** Ensure the upkeep of the data warehousing, data cleaning, data staging, data processing and data architecture. During this stage, the raw data is transformed into a format that can be utilised by the organisation.
3. **Process:** The procedure consists of data mining, clustering and classification, data modelling and data summarization. The produced data is then analysed by data scientists, who look for patterns, ranges and biases in the data to assess how beneficial it will be for predictive analysis.
4. **Analyse:** Conduct several types of analyses, including exploratory and confirmatory analyses, predictive analyses, regression analyses, text mining and qualitative analyses. This phase is the most significant part of the lifespan. At this point in the process, it is time to execute the different analytics on the data.
5. **Communicate:** Communicate with Data Reporting, Data Visualization, Business Intelligence and Decision Making. The last phase in the process is analysts putting the results of the studies into formats that are simple to read, such as charts, graphs and reports.

What is data science used for?

There are four primary approaches to data analysis that are supported by data science:

1. Descriptive analysis

The purpose of descriptive analysis is to investigate the data in order to acquire an understanding of what has occurred or what is occurring in the data environment. It

is distinguished by the use of data visualisations including pie charts, bar charts, line graphs, tables and narratives that are created automatically. A service that books flights, for instance, may keep a record of information such as the daily total of tickets purchased. A descriptive study will show that this service has seen booking spikes, booking slumps and high-performing months in the past.

2. Diagnostic analysis

An in-depth investigation or a comprehensive review of the data can help diagnostic analysts better comprehend the reasons behind an event. Methods like drill-down, data discovery, data mining and correlations are all examples of what make up this type of analysis. On any given data set, multiple data operations and transformations may be carried out in order to uncover one-of-a-kind patterns using any of these methods. For instance, the flight service may zero in on a particularly well-performing month in order to gain a better understanding of the booking surge. This might lead to the revelation that a large number of clients go to a specific city on a regular basis to attend a sporting event.

3. Predictive analysis

The goal of predictive analysis is to provide accurate projections about data patterns that may emerge in the future by making use of past data. Techniques like machine learning, forecasting, pattern matching and predictive modelling are some examples of what are included in this category. In each of these methods, computers are taught to reverse engineer causality connections in the data. For instance, the flight service team might use data science to predict flight booking patterns for the following year at the beginning of each year. This could be done by using historical flight booking data. It's possible that the computer programme or algorithm may analyse the previous data and forecast an increase in bookings for particular locations in May. Due to the fact that the organisation had foreseen the future travel needs of their customers, they were able to begin targeted advertising for those places as early as February.

4. Prescriptive analysis

The next degree of accuracy in data prediction is achieved through the use of prescriptive analytics. Not only does it forecast what will most likely occur, but it also recommends the best course of action to take in reaction to that conclusion. It is able to do an analysis of the potential repercussions that may result from the various options and make recommendations for the most effective next step. It employs techniques from the field of machine learning such as graph analysis, simulation, complicated event processing, neural networks and recommendation engines.

To continue with the example of flight bookings, prescriptive analysis can investigate previously run marketing initiatives in order to make the most of the anticipated increase in bookings. A data scientist may make projections about how many bookings will result from certain levels of marketing spend distributed across a variety of marketing channels. The corporation that books flights might make more informed judgements about their marketing strategies with the help of these data projections.

What is the data science process?

Often, a business challenge serves as the impetus for the beginning of the data

Notes

science process. A data scientist will collaborate with various business stakeholders to gain an understanding of the requirements of the organisation. When the problem has been stated, the data scientist may attempt to address it utilising the OSEMN data science method, which includes the following steps:

O – Obtain data

The data may already exist, be newly gathered, or come from a repository that is accessible over the internet and may be downloaded by the user. Data scientists are able to gather information from a variety of sources, including internal and external databases, customer relationship management (CRM) software, web server logs, social media and reliable third-party sources that they may purchase.

S – Scrub data

The act of standardising the data such that it conforms to a format that has been defined in advance is known as “data cleaning” or “data scrubbing.” The processing of missing data, the correction of data inaccuracies and the elimination of any data outliers are all included in this process. The following are some instances of data scrubbing:

- Converting all the date values into a format that is consistent and universal.
- Correcting any misspellings or adding missing or extra spaces.
- Correcting mathematical errors or deleting commas from very huge numbers.

E – Explore data

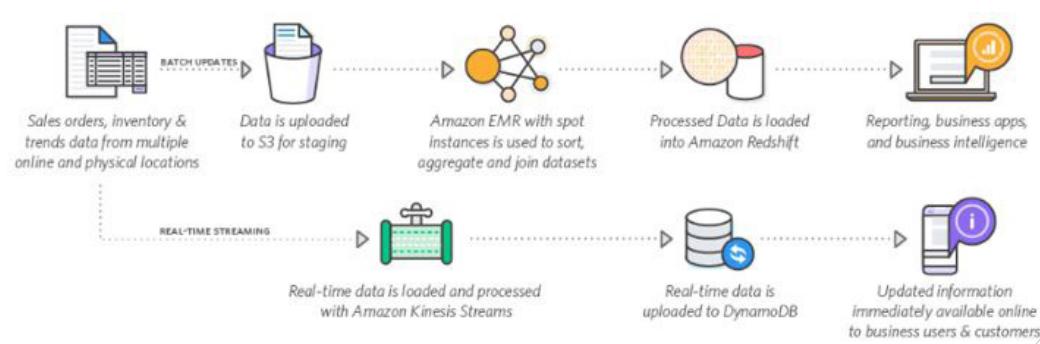
The first step in the data modelling process is called data exploration and it consists of doing basic analyses of the data. Data scientists begin to develop a fundamental comprehension of the data by employing descriptive statistics and other tools for data visualisation. After that, they investigate the data to look for intriguing patterns that might be investigated further or put into action.

M - Model data

In order to get more in-depth insights, forecast results and recommend the most effective course of action, software and machine learning algorithms are utilised. The training data set is used to teach machine learning algorithms such as association, classification and clustering. In order to establish how accurate the results are, the model might be compared to some specified test data. There are several ways in which the data model may be adjusted to get better results.

N – Interpret results

Data scientists collaborate with other roles within an organisation, such as analysts and businesspeople, to turn data insights into action. They show patterns and forecasts using diagrams, graphs and charts that they develop themselves. A concise presentation of the data aids stakeholders in both understanding and effectively implementing the results.



Notes

What are the data science techniques?

The process of data science is often carried out by experts using various kinds of computing technologies. The following are the primary methods that data scientists employ:

a) Classification

The process of organising data into distinct groups or categories is known as classification. It is possible to teach computers to recognise and organise data. In order to develop decision algorithms in a computer that can swiftly analyse and categorise the data, known data sets are employed as building blocks. For example:

- Classify items according to whether they are popular or not.
- Sift applications for insurance into high risk and low risk categories.
- Sort social media comments into favourable, negative, or neutral.

The process of data science is often carried out by experts using various kinds of computing technologies.

b) Regression

Finding a connection between two data items that at first glance appear to have no bearing on one another is the goal of the statistical technique known as regression. Typically, the link is modelled based on a mathematical formula and the resulting model is either a graph or a set of curves. Regression is used to forecast the value of the other data point when it is known that the value of the first data point. For example:

- The rate at which illnesses can be transmitted through the air.
- The connection between the quantity of workers and the level of satisfaction experienced by customers.
- The correlation that exists between the number of fire stations present in a given area and the total number of people who sustain injuries as a result of fires.

c) Clustering

Clustering is a process that involves grouping data that is closely linked together for the purpose of searching for patterns and outliers. The data cannot be correctly sorted into predetermined groups, which is one of the main differences between clustering and sorting. As a result, the data are organised into the correlations that are the most plausible. Clustering allows for the discovery of new patterns and connections between things. For example: ·

Notes

- In order to provide superior customer service, you should categorise clients according to their purchasing patterns.
- Organise network traffic into groups in order to recognise everyday use patterns and locate an assault on the network more quickly.
- Organise the articles into a number of distinct news categories and then utilise this information to locate stuff that is not legitimate news.

The Fundamental Idea that Underlies Various Data Science Practises

Although the specifics differ, the following general ideas underlie each of these approaches:

- Train a machine to sort data based on a data set that is already known to it. For illustration purposes, the computer is provided with sample keywords along with their respective sort values. Positive words include "Joy," whereas negative words include "Hate."
- You should provide the machine with data that is unfamiliar to it and then let the gadget sort the dataset on its own.
- Accommodate for any errors in the results and deal with the associated probability factors.

Data Science Tools

1. Apache Spark

According to its proponents, Apache Spark is a data processing and analytics engine that is open source and can handle massive volumes of data (upwards of several petabytes). Since its inception in 2009, a major increase in the use of the Spark platform has been spurred by the capacity of Spark to swiftly process data. This development has contributed to the Spark project becoming one of the largest open-source communities among big data technologies.

Spark is ideally suited for use in applications that require continuous intelligence and are powered by the near-real-time processing of streaming data because to its speed. Spark, on the other hand, is a general-purpose distributed processing engine that is equally suitable for extract, transform and load purposes as it is for other SQL batch processes. Initially, Spark was promoted as a speedier alternative to the MapReduce engine for batch processing in Hadoop clusters. This is true.

Spark is still often used in conjunction with Hadoop, although it is also capable of operating alone against a variety of file systems and data repositories. It makes it simpler for data scientists to swiftly put the platform to use by providing a comprehensive collection of developer libraries and application programming interfaces (APIs), which includes support for important programming languages and a library devoted to machine learning.

2. D3.js

D3.js is a JavaScript framework that may be used in a web browser to generate individualised representations of data. It is one of the open-source tools available. Instead of using its own graphical vocabulary, it makes use of web standards such

as HTML, Scalable Vector Graphics and CSS. D3, which is an abbreviation for Data-Driven Documents, is the common name for this format. The people who created D3 describe it as a tool that is both dynamic and versatile and that in order to build visual representations of data, it takes a minimum amount of work.

Visualization designers may use D3.js to link data to documents by way of the Document Object Model (DOM) and then utilise DOM manipulation functions to make data-driven changes to the pages they are working with. It was first made available to the public in 2011 and it enables features like as interactivity, animation, annotation and quantitative analysis. It may be used to construct a variety of different sorts of data visualisations.

On the other hand, D3 contains more than 30 modules and 1,000 different visualisation approaches, making it a challenging programme to master. In addition, a significant portion of data scientists are not proficient in JavaScript. As a consequence of this, they could feel more at ease using a commercial visualisation tool, like as Tableau. As a consequence of this, data visualisation developers and specialists who are also part of data science teams are more likely to utilise D3.

3. IBM SPSS

IBM's Statistical Package for Social Sciences (SPSS) is a set of software programmes designed to manage and analyse complicated statistical data. It consists of two primary products: SPSS Statistics, which is a tool for statistical analysis, data visualisation and reporting; and SPSS Modeler, which is a platform for data science and predictive analytics with a drag-and-drop user interface and machine learning capabilities. Both of these products can be found on the company's website.

Users of SPSS Statistics are able to, among other things, explain correlations between variables, generate clusters of data points, discover trends and make predictions. SPSS Statistics covers every stage of the analytics process, from planning through model deployment. It has a menu-driven user interface, its own command syntax and the capability to incorporate R and Python extensions, in addition to tools for automating operations and import-export linkages to SPSS Modeler. It can access common structured data formats.

The software for statistical analysis was initially released by SPSS Inc. in 1968 under the name Statistical Package for the Social Sciences. In 2009, IBM purchased SPSS Inc. along with the predictive modelling platform that SPSS had previously purchased. Both of these pieces of software are now owned by IBM. Although though the programme is part of a product family that is formally named as IBM SPSS, it is most commonly referred to just as SPSS.

4. Julia

In addition to being used for numerical computation, the open-source programming language Julia is also put to use in machine learning and a variety of other data science-related applications. One of the primary objectives of Julia is to eliminate the necessity of writing computer programmes in one language and then converting those programmes to run in another language.

Julia accomplishes this goal by combining the benefits of a high-level dynamic language with the efficiency of a statically typed language, such as C or Java, in

Notes

terms of its performance. Users are not required to create data types in programmes; nevertheless, an option gives them the ability to do so. The utilisation of a multiple dispatch technique during runtime is another factor that contributes to the acceleration of execution speed.

Julia 1.0 was released in 2018, nine years after development on the language first started; the most recent version is 1.8.4 and a beta version of the Julia 1.9 upgrade is now available for testing. The documentation for Julia states that new users “may find that Julia’s performance is unintuitive at first.” However, “once you understand how Julia works, it’s easy to write code that’s nearly as fast as C,” the documentation adds. This is due to the fact that Julia’s compiler is different from the interpreters used in other data science languages, such as Python and R.

5. Jupyter Notebook

Jupyter Notebook is a web tool that is open source and it enables users to collaborate interactively on projects with other users, including data scientists, data engineers, mathematicians and academics. It is a tool for creating computational notebooks that may be used to generate, modify and exchange code in addition to other material, such as descriptive prose, photographs and other data. Users of Jupyter, for instance, have the ability to incorporate software code, computations, comments, data visualisations and rich media representations of computation results into a single document referred to as a notebook. This notebook can then be shared with other individuals and edited by those individuals.

According to the documentation for Jupyter Notebook, as a consequence of this, notebooks “may serve as a comprehensive computational record” of interactive sessions held by members of data science teams with one another. Documents in the notebook are stored as JSON files, which provide version control functions. Users who do not have Jupyter installed on their computers will be able to see notebooks thanks to a service known as the Notebook Viewer, which renders the notebooks in the form of static web pages.

The computer language Python is where Jupyter Notebook got its start; before its separation from the IPython Interactive Toolkit Open Source Project in 2014, it had been a component of that project since its inception. Jupyter got its name from a rather loose mix of Julia, Python and R. In addition to supporting those three languages, Jupyter offers modular kernels for dozens of other programming languages. JupyterLab is an updated web-based user interface that is included in the open source project. Compared to the first UI, JupyterLab is more adaptable and extendable.

6. Keras

Keras is a programming interface that simplifies access to and utilisation of the TensorFlow machine learning platform for data scientists. Keras was developed by Google. It is an open source deep learning API and framework written in Python that works on top of TensorFlow and is now integrated into that platform. Both of these components were originally developed by Google. Keras once supported a number of different back ends, but beginning with the 2.4.0 release of TensorFlow in June 2020, it became solely dependent on that library.

Keras was developed to be a high-level application programming interface (API) that enables easy and rapid experimentation while requiring less code than other solutions for deep learning. The documentation for Keras uses the phrase “high iteration velocity” to describe the objective, which is to “accelerate the construction of machine learning models,” and in particular, deep learning neural networks, using a development process with “high iteration velocity.”

The Keras framework comes with two types of user interfaces: a sequential one for building relatively straightforward linear stacks of layers with inputs and outputs and a functional one for building more complex graphs of layers or writing deep learning models from scratch. The sequential interface is used for creating relatively simple linear stacks of layers with inputs and outputs. Keras models may be deployed across numerous platforms, including web browsers as well as Android and iOS mobile devices. They can also operate on central processing units (CPUs) or graphics processing units (GPUs).

7. Matlab

Matlab is a high-level programming language and analytics environment for numerical computation, mathematical modelling and data visualisation. Since 1984, the software vendor MathWorks has been the company responsible for developing and selling Matlab. Data analysis, algorithm development and the creation of embedded systems for wireless communications, industrial control, signal processing and other applications are the primary uses for this software. It is typically used in conjunction with a companion Simulink tool that provides model-based design and simulation capabilities.

Matlab is not as extensively used in data science applications as languages such as Python, R, or Julia; nonetheless, it does enable machine learning and deep learning, predictive modelling, big data analytics, computer vision and other work that is performed by data scientists. The data kinds and high-level functionalities that are included into the platform are aimed to expedite exploratory data analysis as well as data preparation in applications that deal with analytics.

Matlab, which is an abbreviation that stands for “matrix laboratory,” is a programme that can be learned and utilised with relative ease. It has a number of prebuilt applications, but it also gives users the ability to construct their own. It also features a library of add-on toolboxes that contain software that is particular to a given field, as well as hundreds of built-in capabilities, one of which is the capability to visualise data in both 2D and 3D plots.

8. Matplotlib

Matplotlib is a plotting library written in Python that is open source and is utilised in analytics applications for reading, importing and visualising data. Matplotlib is a library that can be used in Python scripts, the Python and IPython shells, Jupyter Notebook, web application servers and a variety of GUI toolkits. It enables users, including data scientists and other users, to build data visualisations that are static, animated and interactive.

The huge code base of the library can be difficult to grasp, although it is arranged in a hierarchical manner that is meant to enable users to generate visualisations

Notes

primarily using high-level commands. Despite this, mastering the library's code base can be problematic. The most important component in the hierarchy is called pyplot and it is a module that offers a "state-machine environment" as well as a collection of straightforward charting routines that are comparable to those found in Matlab.

Matplotlib was first made available to the public in 2003. It features an object-oriented interface that may be used in conjunction with pyplot or on its own. Moreover, it enables low-level instructions for data charting that is more complicated. The library is largely geared towards the production of 2D visualisations, however it also include an add-on toolset with capabilities for 3D plotting.

9. NumPy

NumPy is an acronym that stands for Numerical Python. It is the name of an open-source Python library that is utilised extensively in applications relating to scientific computing, engineering, data science and machine learning. The library contains objects that are multidimensional arrays and algorithms for processing those arrays in order to enable a variety of mathematical and logical operations. In addition to that, it provides linear algebra, the production of random numbers and other activities.

The N-dimensional array, often known as an ndarray, is one of the fundamental building blocks of NumPy. This component represents a collection of objects that are all the same size and type. The structure of the data elements contained within an array can be described by an associated data-type object. Several ndarrays are able to share the same data and any modifications that are done to the data in one may be viewed in another.

In 2006, components of two older libraries were combined and modified to produce NumPy, which was then released to the public. The website for NumPy refers to it as "the worldwide standard for working with numerical data in Python," and it is widely regarded as one of the most helpful libraries for Python due to the numerous built-in functions that it has. Also, it is renowned for its speed, which is in part attributable to the utilisation of C code that has been optimised at its core. In addition, several other Python libraries are constructed on top of the NumPy foundation.

10. Pandas

Pandas is yet another well-known open-source Python library. Its primary purpose is to do data manipulation and analysis. Created on top of NumPy, it has two core data structures: the Series one-dimensional array and the DataFrame, a two-dimensional structure for data processing with integrated indexing. Both of these structures may be accessed using the Python interpreter. Both can take data from NumPy ndarrays as well as other inputs, but a DataFrame has the additional capability of including numerous Series objects.

Pandas was first released in 2008 and features built-in data visualisation capabilities, exploratory data analysis methods and support for a variety of file types and languages including CSV, SQL, HTML and JSON. Pandas was also one of the first open-source programming languages. According to the website for pandas, it also has the capabilities of intelligent data alignment, integrated management of missing data, flexible reshaping and pivoting of data sets, data aggregation and transformation and the capability to swiftly merge and combine data sets.

The developers of pandas have stated that their goal is to make it “the fundamental high-level building block for doing practical, real-world data analysis in Python.” Key code paths in pandas are written in C or the Cython superset of Python in order to optimise its performance and the library can be used with a variety of different types of analytical and statistical data, including tabular, time series and labelled matrix data sets.

Python is the computer language that is used the most frequently in the fields of data science and machine learning. It is also one of the most popular languages in general. The website for the Python open source project refers to it as “an interpreted, object-oriented, high-level programming language with dynamic semantics.” Moreover, it has built-in data structures and features for dynamic typing and binding. The website also highlights Python’s straightforward syntax, stating that the programming language is simple to learn and that the fact that it places a focus on readability lowers the cost of software maintenance.

The general-purpose programming language is applicable to a broad variety of endeavours, including as the analysis of data, the display of data, artificial intelligence, the processing of natural language and the automation of robotic processes. Python allows programmers to construct programmes for desktop computers, mobile devices and the web. It supports not just object-oriented programming but also procedural, functional and other styles of programming, in addition to extensions written in C or C++.

11. Python

Python is utilised not just by professionals within the realm of computer, such as data scientists, network engineers and programmers, but also by workers outside of the realm of computing, such as accountants, mathematicians and scientists, who are frequently drawn to its user-friendly character. Python 2.x and 3.x are both versions of the language that are fit for production use, despite the fact that support for the 2.x line will expire in 2020.

12. PyTorch

PyTorch is a deep learning framework that is open source and is used to develop and train deep learning models that are based on neural networks. It is lauded by its advocates for facilitating quick and flexible experimentation as well as a seamless transition to production deployment. In comparison to Torch, an earlier machine learning framework built on the Lua programming language, the Python library was developed to have a more intuitive interface and be simpler to use. PyTorch, according to the people who developed it, also offers greater flexibility and speed than Torch does.

PyTorch was made available to the public for the first time in 2017 and it uses arraylike tensors to represent model inputs, outputs and parameters. PyTorch has built-in support for running models on GPUs and its tensors are comparable to the multidimensional arrays that are supported by NumPy. But, PyTorch has an additional advantage. For the sake of processing in PyTorch, NumPy arrays may be transformed into tensors and the reverse is also possible.

The library features a variety of functions and methods, such as a package for automated differentiation known as `torch.autograd`, a module for creating neural

Notes

networks, a tool for delivering PyTorch models known as TorchServe and deployment support for iOS and Android devices. PyTorch provides a C++ application programming interface (API) in addition to its core Python API. This C++ API may either be used independently as a front-end interface or to develop add-ons for Python programmes.

13. R

R is a free and open-source platform that may be used for statistical computation and graphical application development. It can also be used for data processing, analysis and visualisation. R is one of the most popular languages for data science and advanced analytics since it is used by a large number of data scientists, university researchers and statisticians. These individuals use R to retrieve, cleanse, analyse and display data.

The open source project is supported by The R Foundation and thousands of user-created packages with libraries of code that enhance R's functionality are available. One prominent example of this is ggplot2, a package for the creation of graphics that is included in a collection of R-based data science tools known as tidyverse. In addition, integrated development environments and commercial code libraries are also available for R from a variety of different suppliers.

R, much like Python, is an interpreted programming language and it has a well-deserved reputation for being easy to pick up. It was developed in the 1990s as an alternative version of S, a statistical programming language that was established in the 1970s; the name R is both a play on the letter S and a reference to the first letter of the names of its two developers. R was initially developed as an alternate version of S.

14. SAS

The Statistical Analysis System (SAS) is an integrated software package that may be used for statistical analysis, advanced analytics, business intelligence and data management. Users are able to integrate, cleanse, prepare and alter data using the platform, which was developed and is provided by the software vendor SAS Institute Inc. Users are also able to analyse the data using a variety of statistical and data science approaches. SAS is versatile software that may be utilised for a variety of purposes, including but not limited to fundamental business intelligence and data visualisation, risk management, operational analytics, data mining, predictive analytics and machine learning.

The creation of SAS began in 1966 at North Carolina State University. The technology's application began to increase in the early 1970s and in 1976, SAS Institute was established as a separate corporation. The acronym SAS stands for statistical analysis system, which was the original target audience for the programme when it was first developed. But over the course of time, it was developed to incorporate a comprehensive set of functions and it eventually became one of the analytics suites that is used most frequently in commercial companies as well as academic institutions. SAS Viya, a cloud-based version of the platform that was introduced in 2016 and will be modified to be cloud-native in 2020, is now receiving the majority of the attention that is being directed into development and marketing.

1.1.2 Benefits and Uses of Data Science

Importance of Data Science for Business

The use of data science within businesses makes it possible to analyse and monitor performance criteria, which in turn promotes the development and expansion of the company. The models used in data science may reproduce a wide variety of processes by making use of data that already exists. Because of this, firms are able to prepare for the best possible outcomes. The following are some examples of the relevance of data science in business:

1. **Data science for business decision-making:** The use of data science to decision-making in business allows businesses to determine the effectiveness of their operations by basing their reporting on data that is both accurate and up to date. For the purpose of assisting businesses in making educated judgements on significant planning, "business intelligence" gives crucial data on the company's recent and historical productivity as well as future estimates, projected demands, buying patterns and other relevant topics. The goal of the business analytics teams is to ensure that the company receives real-time, improved reports so that it can make better use of the data that is available to run the business more efficiently.
2. **Making quality products:** Companies require data in order to optimise product development in a way that satisfies the demands and expectations of their customers in order to provide great products. Companies are able to create superior goods by doing analysis of their customer data.
3. **Effective business management:** By the use of data science, both small and large organisations are able to effectively manage their operations and further improve themselves. Companies are now able to forecast the success of their strategy by utilising data science.
4. **Forecasting using predictive analysis:** In business, one of the most important applications of data science is forecasting, which may be accomplished via predictive analysis. In order to improve their data mining abilities, businesses make use of various analytical tools and technologies. Through predictive analysis, companies may get insights into factors that may have an impact on their operations and then take the steps necessary to address those factors.
5. **Leveraging data for business decisions:** Using data in business choices Without conducting surveys, organisations would be forced to make decisions that were not in their best interest, which would result in financial losses.
6. **Evaluating business resolutions:** Businesses can make correct business choices fast by anticipating future events and trends. These projections may be used in the evaluation of business resolutions. The company should have a comprehensive understanding of how the resolutions that are put into action will effect their growth and performance.
7. **Fraud and risk management:** Due to their level of experience, data scientists are able to spot data that sticks out from the rest, which is useful in the prevention and control of fraud and risk. After that, they will be able to construct a network, a route and data-driven methods that anticipate fraud.

Notes

- 8. Recruiting automation:** In this day of fierce competition for top performers, many companies have realised that the traditional procedure for hiring just isn't as successful as it used to be. This realisation has led to the rise of recruiting automation. These companies have set for themselves the goal of achieving greater success in a shorter amount of time, more frequently and with less resources than is required to accomplish the goals they have set for themselves.

The Role of Data Science in Business

Collect information pertaining to the customers: Information on a variety of areas, such as a customer's hobbies, demographics, goals and other related topics, may be gleaned from customer data. An understanding of data science makes it easy to comprehend the many data possibilities that are available to the clients.

1. Increase your level of security

Data science provides an opportunity for businesses to improve their internal security and better protect vital data. With the assistance of both computer algorithms and human judgement, companies may get closer to achieving a greater level of data protection and efficiency in their data utilisation.

2. Reporting on the company's finances internally

The company's finance team has the ability to employ data science to generate reports, formulate projections and investigate financial patterns. Because of the wealth of information that can be gleaned from each of these financial analyses, you will be in a position to make educated choices regarding the direction of your business.

3. Efficient production

One such use of data science that the company may employ is to determine where bottlenecks exist in the production processes. Manufacturing machines are responsible for collecting large volumes of data from the various stages of production. By embracing data science in order to become more efficient, businesses have a better chance of cutting costs while simultaneously boosting output.

4. An analysis and forecast of upcoming industry trends

By gathering and examining data over a wider scope, the organisation has a better chance of identifying emerging patterns in the market. Keeping up with the habits of one's target market may be an effective way to get an advantage over one's competitors when it comes to making business decisions that will put one's company ahead of the competition.

Benefits of Data Science for Business

The primary goal of data science is to educate company owners about data deviation, in addition to topics such as the pace of business growth, customer statistics and other relevant topics. The application of data science has significantly improved business operations all around the world. The following are a few of the important advantages that data science offers to businesses.

1. New business ideas and improved infiltration:

Data scientists use machine

learning to develop better ways to identify complex business challenges, which helps them come up with new company ideas and enhance infiltration. It is possible that they will discover errors that were overlooked. Data scientists are involved in the reporting of advances in their respective industries, resource-based expenditures, profit estimates and improving the efficiency of the business plan by providing well-informed objectives.

2. **Betterment of products and services:** Improvements to both products and services A company's primary goal should be to provide clients with enhanced offerings that are worthy of their repeat business. The level of happiness experienced by the consumer will determine everything, including earnings and income. Data science contributes to the process of developing consumer goods by analysing feedback from customers, investigating current and future market trends, contrasting two competing products and selecting the superior of the two options based on its capacity to attract and retain customers over an extended period of time.
3. **Malware prevention and improvements:** By studying user data and gaining an understanding of market and customer behaviour, the company will have a lot more narrow viewpoint that is free from ideological or special prejudices. This will allow the company to better avoid and enhance malware. As a consequence of this, the firm will be in a position to recognise any issues or concentrate on the optimisations that are necessary to grow the business.
4. **Attractive campaigning:** The firm may prepare to engage in ads, programmes and campaigns and boost the impact of each investment by having data on user behaviour. This will make the company more appealing to potential customers.
5. **Reduce potential dangers:** efficient data science and analytics make it possible to combat fraudulent activity in real time and improve overall safety. It is possible for it to aid firms in detecting other abnormalities that may undermine their security, in addition to spotting prospective cyberattacks.
6. **Optimization of the warehouse:** Warehouse management that is informed by data science may dramatically improve customer service by cutting down on both excess inventory and urgent orders while simultaneously boosting inventory turnover rates. This results in an optimised warehouse.

How to Use Data Science for Business Growth

The data-driven economy serves as the support structure for today's businesses. The collecting and processing of data in the right way helps a company expand, whereas using inaccurate data drives up operating costs. The following are a few of the many ways that data science contributes to the expansion of enterprises.

- **Empowered to create improved resolutions:** When organisations employ data-driven decision-making, they focus on information that may assist them in making better judgements. This information can be found in a variety of sources. Making choices based on data not only increases corporate responsibility and transparency, but also increases employee engagement. The information may be used to address issues relating to the development of the business, as well as financial problems, sales and marketing and the quality of the service.
- **Define objectives based on trends:** When trends are recognised via the use of

Notes

data obtained from a variety of search engines, the performance of the institution is better, customers are engaged with the company in a more productive manner and finally, profitability is raised.

- **Educating the team:** When data science is implemented, the organisation may quickly discover insights that are helpful to its staff. This may be accomplished through training. In addition, this data might be used to publish information on websites or in permanent records, both of which would be accessible at any time to members of the workforce.
- **Automate processes:** Automation may save time and money by automatically extracting, generating, or interpreting material. This can be done in a process that is automated. It is becoming an increasingly important skill to have in this day and age of large data warehouses because the data stored therein lacks any inherent order.
- **Construct superior goods:** Data science can be applied to business in one of two ways: either by personalising a good or service so that it can be used by a specific customer, or by providing a novel way for customers to make use of the good or service. Either way, the goal is to produce superior goods that can be sold to the target market.
- **Evaluating opportunities:** The data science opportunity assessment enables quickly determining the most valuable data science prospects for the company by identifying trade-offs that must be managed and gaps that must be filled. This is accomplished by identifying gaps in knowledge that must be filled.
- **Identifying and focusing on the target audience:** Determining who your ideal customers are and concentrating your efforts on them the vast majority of businesses have channels through which they collect information on their customers, such as Product Surveys or General Surveys. The information that was gathered is rendered useless if it is not employed in the appropriate manner, such as in the case of demographic data.
- **Right employee selection:** Your hiring staff may be able to make decisions more quickly and accurately with the help of data science. This can be accomplished through data mining, the processing of internal applications and resumes and even the utilisation of sophisticated data-driven aptitude tests. The HR department can sift through all of the information that is collected from the many job portals and database providers in order to locate applicants who are the most suitable for the organisation. It helps save time while also selecting the most qualified candidates.

Five Trends Influencing the Use of Data Science in Business

It is impossible to ignore the developments in data and analytics that are reflective of the changes in the corporate world, the market and technology. These tendencies encourage new growth, efficiency, resilience and innovation, which helps with the prioritising of investments.

1. **Data science for business growth and automation:** The use of data science to the expansion and automation of business processes Data science provides several alternatives for the improvement of business procedures. Businesses have the ability to apply the studied data in their production in order to remove downsides, improve

resource efficiency and select the appropriate quality. With the use of data science, manufacturers are able to alleviate the issues that arise throughout production. In turn, this impacts how activities related to product quality, supply and delivery are carried out.

2. **Intrinsic concepts:** When it comes to promotional efforts, the creation of new products, or the selection of content, adopting data science may reduce a significant number of the limits. The use of data analytics enables a more complete perspective of the customers, as well as a better comprehension of what their requirements are and how best to fulfil those requirements.
3. **Solution for artificial intelligence and big data in the cloud:** The collection, analysis, purification, organisation and storage of the huge volume of data is a considerable challenge. As a direct consequence of this, businesses are increasingly embracing cloud-based solutions. Market expansion will be driven in part by both the rising demand for intelligent systems and the expanding adoption of cloud-based solutions across a number of end-user sectors.
4. **Boost performance and competition:** Machine learning algorithms are able to find patterns and insights in data, which can be used for making more accurate decisions or predictions, the classification of images and object recognition, the detection of fraudulent, unique and specific information and many other applications. This can boost performance and competition.
5. **Data science and blockchain:** Since the blockchain is a decentralised ledger, data scientists are able to make the optimal judgement straight from their devices. By utilising decentralised ledgers, it is possible to simplify the process of managing huge volumes of data.

Strategies to Improve Your Business Using Data Science

For a company's continued success, astute business tactics are constantly required. The following points provide an explanation of how to apply data science in a business setting.

- **Data mining and analysis:** In data mining, large data sets are sorted in order to uncover patterns and correlations. These patterns and associations may then be utilised in data analysis to aid in the resolution of business difficulties. Businesses are able to better anticipate future trends and make decisions based on more accurate information when they employ the methodologies and technology of data mining.
- **The selection of the final choice:** The optimal and most effective decision should be picked from among the analytical possibilities. This last decision will ultimately determine the success of the organisation.
- **Solution for artificial intelligence and big data in the cloud:** The company's data bank is kept up to date and error-free by data scientists who actuarially chose useful data, which contributes to the company's control of the information. The company consults this data bank for various purposes when the need arises.
- **Safety and security:** Because the safety and security of data banks is such an important concern, appropriate protections are required to guarantee that sensitive

Notes

firm information does not end up in the hands of dishonest business rivals or criminals.

- **Automation of processes:** Automation depends on mistake-free data instructions and assists organisations with time management, actuarial selection and cost reduction. Automation also reduces the risk of human error.
- **Providing training to the members of the work team:** Providing training to the members of the work team on how to utilise and profit from the data bank is always helpful and helps in the accomplishment of their jobs.

1.1.3 Role of Data Scientist

Data scientists are professionals in data analysis and possess the technical abilities necessary to solve difficult challenges. They collect, examine and evaluate massive volumes of data while dealing with a range of ideas linked to computer science, mathematics and statistics. They have a responsibility to provide viewpoints that go beyond the study of statistical data. It is possible to find work as a data scientist in a variety of fields, including banking, consulting, manufacturing, pharmaceuticals, government and education, among others.

A Data Scientist's Duties and Obligations in Today's World

- **Management:** Although the Data Scientist does not play a significant managerial role, he does provide support for the construction of a base of futuristic and technical abilities within the Data and Analytics field. This is done in order to assist a variety of planned and ongoing data analytics projects.
- **Analytics:** The Data Scientist is a scientific profession that entails the planning, implementation and evaluation of high-level statistical models and strategies for the purpose of application in the most difficult problems facing the company. Data Scientists are responsible for developing economic and statistical models to solve a variety of issues, such as projections, classification, clustering, pattern analysis, sampling, simulations and so on.
- **Strategy and Design:** The Data Scientist plays an important part in the development of innovative strategies to understand the business's consumer trends and management, as well as ways to find solutions to challenging business problems, such as the optimisation of product fulfilment and entire profit. This role is essential to the advancement of the field of data science.
- **Collaboration:** The function of the Data Scientist is not a lonely role and in this position, he collaborates with other outstanding data scientists to convey difficulties and discoveries to essential stakeholders in an effort to increase drive company performance and decision-making.
- **Knowledge:** The Data Scientist is also responsible for taking the lead in exploring a variety of technologies and tools with the goal of developing novel data-driven insights for the company in the quickest and most agile manner that is practically possible. In this scenario, the Data Scientist takes the initiative to evaluate and implement new and improved data science methodologies for the company, which he then presents to top management for their approval.

- **Additional Responsibilities:** A Data Scientist is also responsible for doing activities that are connected to their job as well as tasks that have been delegated to them by the Senior Data Scientist, Head of Data Science, Chief Data Officer, or the Employer.

1.2 Big Data and Data Science

Introduction

Big data is a term used to describe extensive and varied collections of data that are accumulating at ever-faster rates. The “three v’s” of big data are the volume of information, the velocity or speed at which it is generated and gathered and the variety or breadth of the data points that are being covered. All of these aspects are included in the concept of “big data.” Data mining is frequently the source of big data, which then arrives in a variety of formats.

1.2.1 Definition: What is Big Data

The term “Big Data” refers to collections of data that are extremely extensive. We often interact with data that is either megabytes (MB) or gigabytes (GB) in size (movies, codes), but the term “big data” refers to data that is petabytes, or 10^{15} bytes, in size. According to some estimates, over 90 percent of the data used today was created during the last three years.

Sources of Big Data

These data originate from a wide variety of sources, such as:

- **Social networking sites** such as Facebook, Google and LinkedIn, each of which creates a massive quantity of data on a day-to-day basis due to the fact that they have billions of members all over the world.
- **E-commerce website:** Websites such as Amazon, Flipkart and Alibaba create a large quantity of logs, which may be used to determine the purchasing patterns of their customers.
- **Weather Stations:** Every weather station and satellite contributes very large amounts of data, which are then saved and modified in order to provide weather forecasts.
- **Telecom company:** Telecom giants like Airtel and Vodafone examine the tendencies of their users and publish their services in accordance with those findings; in order to do this, they save the information of their millions of customers.
- **Share Market:** Each day's trading activity at stock exchanges throughout the world results in the creation of a massive amount of data.

3V's of Big Data

Volume

Quantity is the most important aspect of big data. data quantities that might reach heights that were previously inconceivable. According to estimations, 2.5 quintillion

Notes

bytes of data are created every single day, which means that by the year 2020, there will be a total of 40 zettabytes of data created. This is a 300-fold increase from the year 2005. As a direct consequence of this, large companies today routinely keep terabytes and even petabytes of data in their storage and on their servers. This information is helpful in designing the operations and future of a firm while also tracking its progress.

Velocity

The way that we think about data has evolved as a result of both the growth of data and the significance that it has taken on. The importance of data in the corporate world was previously underappreciated by us, but as a result of improvements in the methods by which we collect it, we now frequently rely on it. The term “velocity” refers to the rate at which new data is being added to the system. Some of the data we want will be delivered to us in batches, while other pieces will trickle in here and there. Because not all systems analyse incoming data at the same rate, it is essential to refrain from forming assumptions before gathering all of the relevant information.

Variety

The data used to be presented in a singular manner and come from a sole origin. It was once provided in database files such as excel, csv and access files; but, it is now being provided in non-traditional formats through technology such as wearable devices and social media. These formats include video, text, pdf and graphics. Even while we may benefit from this information, interpreting it, managing it and putting it to use requires far more effort and intellectual prowess than we now possess.

How Does Big Data Work?

Finding trends, patterns and correlations among massive volumes of unprocessed data is one of the tasks involved in the analytics of big data. This is done so that judgements may be made based on the data. These approaches use well-known statistical analysis methods to bigger datasets, such as clustering and regression, with the assistance of more modern tools.

1. Data Collection

When it comes to data collecting, every firm takes a somewhat different strategy. Because of advancements in technology, organisations are now able to collect data in both its structured and unstructured forms from a wide number of sources. These sources can include cloud storage, mobile applications, in-store Internet of Things sensors and more.

2. Organise the data

Once the data have been obtained and saved, they need to be adequately organised in order for analytical queries to provide right responses. This is especially important if the data are large and unstructured.

3. Clean Data

To improve the quality of the data and provide more reliable conclusions, it is necessary to clean all of the data, regardless of its quantity. It is vital to eliminate or

account for any redundant or superfluous data and all of the data must be formatted in a suitable manner. Inaccurate conclusions can be drawn from soiled data because of its ability to conceal and fool.

4. Analysis of Data

The process of converting massive volumes of data into a form that can be used takes time. If the data is made public, sophisticated analytics tools may be able to turn massive amounts of data into meaningful insights. Methods such as these for analysing massive amounts of data include:

- Data mining is the process of sifting through vast datasets to locate patterns and relationships. This is done by locating anomalies and building data clusters.
- Predictive analytics examines future forecasts using previous data from a company in order to identify prospective risks and possibilities.
- Deep learning is the process of using several layers of algorithms to discover patterns in even the most complex and abstract data, imitating the way humans learn.

Types of Big Data

Following are the types of Big Data:

1. Structured
2. Unstructured
3. Semi-structured

Structured

The word “structured data” refers to any type of data that can be saved, retrieved and processed in the form of a predetermined format. Talent in computer science has, throughout the course of time, had more success in inventing strategies for working with this sort of data (when the format is well understood in advance), as well as approaches for getting value from the data itself. Yet, in this day and age, we are able to anticipate problems that may arise when the quantity of such data increases to a significant degree; typical quantities are already in the range of several zettabytes.

When one considers these numbers, it is easy to comprehend why the phenomenon is referred to as “Big Data,” as well as the difficulties that are associated with its storage and processing.

Do you know? 10^{21} bytes equal to 1 zettabyte or one billion terabytes forms a zettabyte.

Do you know? Data stored in a relational database management system is one example of a ‘structured’ data.

Examples of Structured Data

A good example of structured data would be the table in a database labelled “Employee.”

Notes

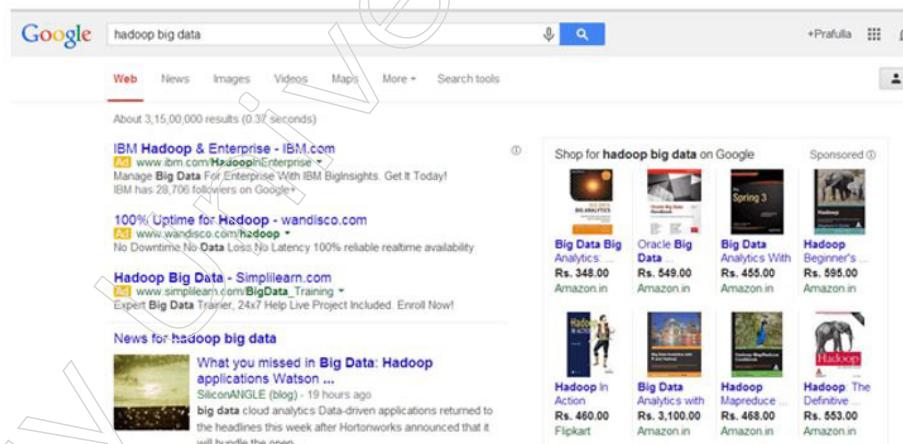
Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured

Unstructured data refers to any data where the form or structure is unclear. This includes most types of data. Unstructured data presents various obstacles in terms of its processing in order to derive value from it. This is in addition to the fact that the amount of the data is enormous. A heterogeneous data source, which may include a mixture of basic text files, photos, videos and other types of media, is a good illustration of unstructured data in its normal form. Due to the fact that the data is stored in an unstructured or raw state, modern businesses have access to a plethora of data; yet, they are unable to extract value from this data since they do not know how to do so.

Examples Of Un-structured Data

The output returned by 'Google Search'



Example of Un-structured Data

Semi-structured

Both organised and unstructured data may be included in semi-structured data. We can look at semi-structured data as if it were structured, but, it is not specified in any way, such as with a table definition in a relational database management system. One example of data that is semi-structured is information that is stored in an XML file.

Examples of Data That is Semi-structured Data

Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
```

```
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Notes

1.2.2 Evolution of Big Data and its Importance



The History of Big Data

The development of data and more specifically big data, has a lengthy and eventful history. During World War II, there were a number of significant technological advances that were created, the most of which were developed for use in military operations. But, as time went on, those developments would eventually become helpful to the business sector and, eventually, the general public, which would result in personal computing becoming an alternative that the average consumer might consider.

1940s to 1989 – Data Warehousing and Personal Desktop Computers

The Electronic Numerical Integrator and Computer, which was the world's first programmable computer, is considered to be the progenitor of electronic storage devices (ENIAC). During World War 2, the United States Army developed it in order to find solutions to numerical issues, such as calculating the range of artillery fire. After that, in the early 1960s, International Business Machines (IBM) launched the first transistorised computer under the name TRADIC. This allowed data centres to transition from serving primarily military goals to serving more broad commercial purposes.

Apple Computers introduced the Lisa, the world's first personal desktop computer with a graphical user interface (GUI), in 1983. Lisa was manufactured by IBM. Throughout the decade of the 1980s, companies such as Apple, Microsoft and IBM

Notes

released a wide variety of personal desktop computers. This resulted in an increase in the number of people purchasing their very own personal computers and having the ability to use them in their own homes for the very first time. So, people of various socioeconomic backgrounds were finally able to access electronic storage.

1989 to 1999 – Emergence of the World Wide Web

Tim Berners-Lee, a British computer scientist, is credited with developing the essential technologies that were necessary to power what is now known as the World Wide Web between the years 1989 and 1993. The HyperText Markup Language, often known as HTML, the Universal Resource Identifier, or URI and Hypertext Transfer Protocol were these online technologies (HTTP). After that, in April of 1993, a decision was taken to liberate the source code for these web technologies so that it may be used by anybody, forever.

Because of this, people, corporations and organisations that had the financial means to pay for an internet service were able to connect to the internet and exchange data with other internet-capable computers when it became possible for them to do so. Because of the increasing number of devices that were able to connect to the internet, there was a significant increase in the volume of information that individuals were able to access and exchange at any one moment.

2000s to 2010s – Controlling Data Volume, Social Media and Cloud Computing

Companies like Amazon, eBay and Google were instrumental in generating massive volumes of online traffic and a mix of organised and unstructured data during the beginning of the 21st century. In addition, Amazon released a beta version of AWS (Amazon Web Services) in the year 2002, making the Amazon.com platform accessible to all software developers. By 2004, more than one hundred apps had been developed for it.

After that, Amazon Web Services (AWS) relaunched in 2006, at which point it began providing a comprehensive selection of cloud infrastructure services, such as the Simple Storage Service (S3) and the Elastic Compute Cloud (EC2). The public introduction of Amazon Web Services (AWS) drew a broad variety of clients, including Dropbox, Netflix and Reddit. These companies were anxious to become cloud-enabled and as a result, they all decided to collaborate with AWS before the year 2010.

Unstructured data also became more widespread as a direct result of the proliferation of social media platforms such as MySpace, Facebook and Twitter. This would involve the transfer of photos and audio files, as well as movies, animated GIFs, status updates and direct messages.

These platforms needed new ways to gather, organise and make sense of this data since such a massive volume of unstructured data was being created at such a rapid rate. This resulted in the development of Hadoop, an open-source framework designed especially for the management of large data sets and the adoption of NoSQL database queries, which made it possible to manage unstructured data (data that does not comply with a relational database model). Both of these developments were made possible as a result of the aforementioned. Because of these new technologies, businesses are now able to collect vast volumes of diverse data, from which they can then derive useful insights, allowing them to make decisions that are better informed.

2010s to now – Optimization Techniques, Mobile Devices and IoT

In the 2010s, the proliferation of mobile devices and the Internet of Things posed the greatest difficulties for big data (Internet of Things). Suddenly, millions of people all over the world were seen going about their daily lives with small, internet-enabled devices in the palm of their hands. These devices gave users the ability to browse the web, engage in wireless communication with other internet-enabled devices and upload data to the cloud. According to a research titled “Data Never Sleeps” that was published by Domo in 2017, we were producing 2.5 quintillion bytes worth of data each and every day.

The proliferation of mobile devices and Internet of Things devices has also led to the collection, organisation and analysis of new kinds of data. The following are some examples:

- Sensor Data (data collected by internet-enabled sensors to provide valuable, real-time insight into the inner workings of a piece of machinery)
- Social Data (publicly available social media data from platforms like Facebook and Twitter)
- Data Relating to Transactions (data from online web stores including receipts, storage records and repeat purchases)
- Information pertinent to health (heart rate monitors, patient records, medical history)

Companies now have access to information that enables them to delve more deeply than ever before into aspects that had not been investigated in the past. These facts include the purchasing behaviour of customers as well as the maintenance frequency and life expectancy of machinery.

The Future of Big Data Solutions

Despite the fact that the future of big data is not totally apparent, there are current trends and projections that can help shed some light on how big data will be managed in the near future. AI (Artificial Intelligence) and automation are by far the most prominent big data technologies. Both of these technologies are easing the process of database administration and big data analysis, making it simpler to translate raw data into useful insights that make sense to important decision makers.

Big data analytics tools can help a company keep up with the rapidly multiplying generation of data, turn meaningless data into powerful information and knowledge, significantly aid in the decision-making process and increase the odds of predicting future outcomes. This is true whether the company is collecting consumer information or conducting business analytics.

Concerns about ethics provide yet another significant roadblock for big data. Legislation passed at the national and state levels over the course of several decades has resulted in a standardisation of the processes by which private businesses and individuals can carry out data collecting and make use of the information they receive. Regulations such as the General Data Protection Regulation (GDPR) are making it abundantly clear that the privacy of customers is one of the highest priorities. As a result, it is imperative that businesses and individuals take data privacy seriously in order to run their operations legally and avoid significant fines. It is possible

Notes

for businesses to maintain a secure environment and safeguard their sensitive customer and employee data by utilising the most recent data collection and analysis technologies, which have been developed with the express purpose of ensuring compliance with such standards.

Importance of Big data



The significance of big data is not contingent on the quantity of data possessed by an organisation. The manner in which the organisation makes use of the information that it has obtained is directly related to its significance. Every organisation does things differently with the data that it has acquired. The more efficiently the firm makes use of its data, the quicker it will expand. The businesses competing in the modern market are required to amass and examine this information because:

1. Cost Savings

When it comes to storing big volumes of data, organisations may reap the benefits of Big Data technologies like Apache Hadoop and Spark, which help them save money in the process. These tools assist firms in determining business practises that are more productive and efficient.

2. Time-Saving

Companies are able to acquire data from a wide variety of sources with the assistance of real-time analytics that run in memory. They are able to swiftly examine data with the assistance of tools such as Hadoop, which enables them to make decisions more quickly that are founded on their discoveries.

3. Get an understanding of the current market circumstances

The study of Big Data enables firms to have a better grasp of the current state of the market. For instance, doing a study of the purchase patterns of customers enables businesses to determine which items are the most popular and, as a result, to make more of those things. This allows businesses to get a competitive advantage over their other rivals.

4. Paying Attention to social media

With the technologies for big data, businesses are able to do sentiment analysis. These things make it possible for them to obtain comments about their firm, or information regarding who is saying what regarding the organisation. Tools for analysing large amounts of data can help businesses enhance their internet presence.

5. Increase the number of new customers you get and the ones you keep.

Clients are an essential resource that are necessary for the success of every organisation. Without establishing a solid foundation of loyal customers, no company can hope to attain lasting success. Nonetheless, despite having a stable consumer base, the businesses are unable to ignore the rivalry that exists in the market. If we are unable to understand what it is that our clients desire, then the success of our businesses will suffer. That will lead to a decrease in the number of customers, which will have a negative impact on the expansion of the firm. Analytics of large amounts of data help organisations recognise patterns and trends connected to their customers. Analysis of the behaviour of customers is the path to a successful business.

6. Provide Marketing Insights while Resolving Issues Faced by Advertisers

All aspects of a business are influenced by analytics performed on big data. It gives businesses the ability to meet the requirements of their customers. The company's product range may be modified with the assistance of big data analytics. It makes certain that marketing initiatives are successful.

7. The engine that powers new product development and innovations

The availability of large amounts of data gives businesses the ability to develop new goods and improve existing ones.

1.2.3 Four V's in Big Data, Drivers of Big Data

Introduction to the 4 V's of Big Data

Finding information that is relevant in today's world is an extremely important task for any company. This type of data is comprised of big data sets that are either unstructured or structured and both types are quite complicated. These data sets are obtained from relevant sources and then they are transported across cloud and on-premise borders. This process is referred to as "web scraping for big data," where "big data" refers to a high volume of material that may be either structured or unstructured and "web scraping" refers to the action of obtaining and sending content from internet sources. The power of high-powered analytics, which leads to intelligent business decisions about cost and time optimisations, product development, marketing campaigns, issue identification and the invention of new company ideas, is largely responsible for the rise in importance of big data. Continue reading and you will learn what "big data" is, how it can be divided down into several dimensions and how scraping for "big data" may assist you in achieving your professional objectives.

The Big Idea Behind Big Data

Big data refers to information that cannot be processed using conventional techniques because it is either too extensive or too complicated. But, in order for information to be valuable, it must first be safeguarded, processed, interpreted and applied in the appropriate manner. The fundamental goal of extracting information from large amounts of data is to obtain new information and patterns that can be examined in order to make more informed strategic and operational decisions. In addition, the analyses of data patterns will assist you in overcoming costly difficulties and will enable you to forecast the behaviour of customers rather than relying on guesswork. One other

Notes

benefit is the ability to exceed one's opponents. Knowledge analysis will be utilised not just by existing rivals but also by new companies in order to compete, innovate and generate income. Also, you are required to stay up. Big data makes it possible to develop new prospects for growth and most companies now have departments whose sole purpose is to gather and analyse data regarding their goods and services, customers and their preferences, rivals and the trends in their respective industries. Each organisation makes an effort to make effective use of this material in order to discover solutions that will enable:

- Cost savings
- Savings in elapsed time
- Do research on the market.
- Manage the reputation of the brand:
- Focus on retaining more of your existing customers
- Resolving advertising and marketing difficulties
- Product development

4 V's of Big Data

Big data is based on the four pillars of volume, variety, velocity and veracity. These four v's constitute the foundation of big data. Let's look at each one in further depth.

a) Volume

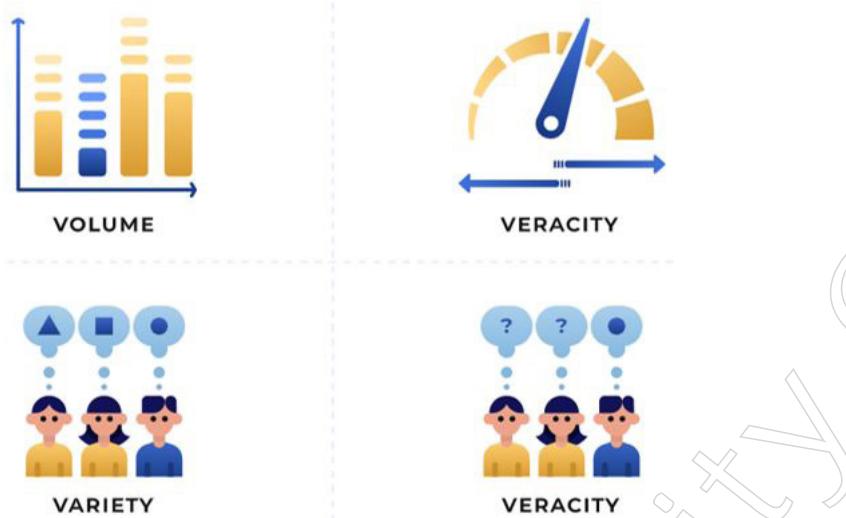
When dealing with a massive amount of information, volume is the most important quality to consider. Big data is measured in petabytes and zettabytes, as opposed to the megabytes, gigabytes, or terabytes that are used to measure ordinary information. In the past, storing material was a difficult challenge. Yet this is now possible because to innovative technologies such as Hadoop and MongoDB. Further mining would not be feasible unless specialised systems for storing and processing information were implemented first. E-mails, social networking platforms, product evaluations and mobile applications are just some of the internet sources that contribute to the massive amounts of data that are collected by businesses. The extent of big data is expected to double every two years, as stated by industry professionals; hence, appropriate data management is going to be absolutely necessary in the years to come.

b) Variety

Due to the fact that huge material can take on a number of forms and comprises both organised and unstructured information, certain processing skills and specialised algorithms are required in order to handle its complexity.

- Demographic numbers, stock insights, financial reports, bank records, product details and other information are all examples of structured content. This material is kept and evaluated with the use of conventional techniques of storage and analysis.
- Unstructured material primarily represents human ideas, sentiments and emotions. This type of information can be documented in the form of video, audio, emails, messages, tweets, status updates, pictures, images, blogs, reviews, recordings and so on. The collecting of unstructured content is accomplished by the application of relevant technologies such as data scraping, which is utilised to

explore websites by going to the greatest possible depth in order to extract useful information for the sake of subsequent research.



c) Velocity

In this day and age, information travels at a breakneck pace and businesses are obligated to process it as quickly as possible. It is important for the information to be created and processed as quickly as possible so that its full potential may be utilised. Although if there are some kinds of material that can retain their usefulness after some amount of time has passed, the vast majority of it demands an immediate response, such as messages on Twitter or postings on Facebook.

d) Veracity

The analysis of the content's veracity should focus on the content's overall quality. When you deal with vast volume, high velocity and such a big diversity, for disclosing genuinely significant figures, you need to employ modern machine learning algorithms. Data with a high level of veracity give information that may be usefully analysed, whereas data with a low level of veracity contain a large number of meaningless numbers that are commonly referred to as noise.

Drivers of Big Data

Big Data is a relatively new concept that evolved during the previous decade as a result of a convergence of growing technological capabilities and evolving business requirements. In the beginning of the 21st century, a number of businesses that put Big Data at the centre of their strategy have achieved a great deal of success in their endeavours. Apple, Amazon, Facebook and Netflix are just a few examples of well-known companies. Big data has swiftly become one of the most highly sought after issues in the sector as a result of a number of business factors that are at the basis of its success and explain why it has done so well. The following list identifies six primary factors that drive business:

1. The increasing use of technology in society
2. The precipitous decline in computing technology
3. Connection achieved via the use of cloud computing

Notes

4. Improvement in one's understanding of data science
5. Social media apps
6. The impending implementation of the Internet of Things (IoT).

First, let us look at each of these business drivers from a high-level perspective. Each of these contributes to the competitive edge that businesses already have by generating additional streams of revenue while simultaneously lowering their operating expenses.

The Digitization of Society

Big Data is focused almost entirely on the end user and is driven by their needs. The majority of the data in the world is produced by customers, who in today's always-connected world, are the source of most of the data. The majority of individuals spend anywhere from four to six hours each day consuming and producing data using a range of electronic devices and (social) applications. Every time you make a tap, swipe, or send a message, fresh information is being added to a database located anywhere in the globe. The amount of data that is being created is beyond comprehension as a direct result of the widespread availability of smartphones. According to the findings of certain research, sixty percent of all data was created during the previous two years. This provides a useful indicator of the pace at which our society has digitised its processes.

The Plummeting of Technology Costs

The cost of the technology required to gather and analyse high varieties of data in vast amounts has decreased significantly in recent years. As the price of data storage and processors continues to fall, it is becoming increasingly feasible for people and smaller organisations to participate in big data projects. Moore's Law, which is frequently referenced, states that the storage density (and, hence, capacity) still doubles every two years. This law applies to the storage capacity. The following picture presents a visual representation of the cost reductions brought about by technological advancements.

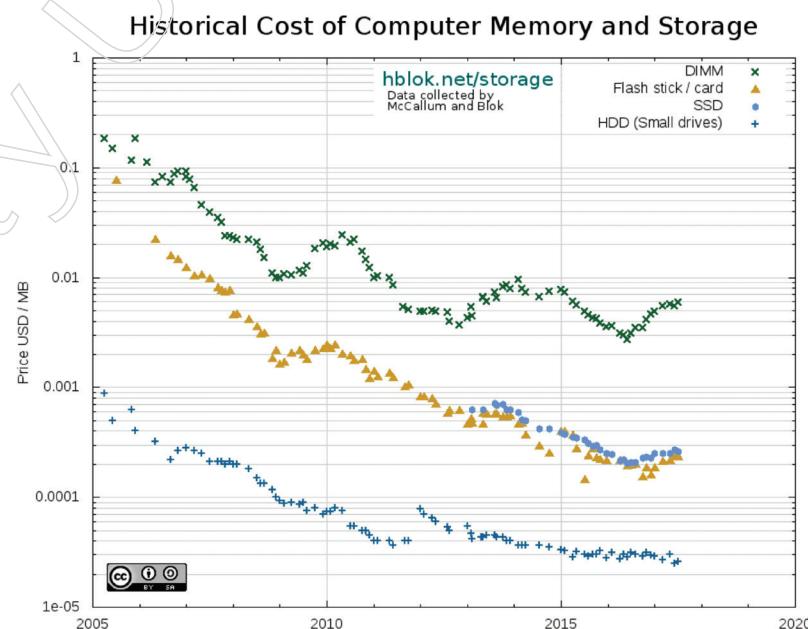


Figure 1: Historical Costs of Computer Memory, reprinted from McCallum and Blok, 2017

The creation of open source Big Data software frameworks has been a significant aspect that has contributed significantly to the affordability of big data, in addition to the precipitous decline in the costs of data storage. Apache Hadoop is the most widely used software framework for distributed storage and processing and it is generally recognised as the industry standard for big data. The widespread availability of these software frameworks in open sources has made it far less expensive for businesses to initiate Big Data projects.

Connectivity Through Cloud Computing

Cloud computing environments, in which data is stored remotely in distributed storage systems, have made it feasible to rapidly scale up or scale down IT infrastructure and to facilitate a pay-as-you-go model. Cloud computing environments also make it possible for users to pay only for the resources that they actually use. This translates to the fact that businesses who wish to process vast volumes of data (and hence have large storage and processing requirements) do not need to invest in significant quantities of IT equipment in order to accomplish their goals. Alternatively, companies may get a licence for the necessary amount of storage and processing capacity and pay only for the resources that they really employ. As a consequence of this, the vast majority of Big Data solutions give their products and services to businesses by using the potential offered by cloud computing.

Increased Knowledge about Data Science

The terms “data science” and “data scientist” have seen a meteoric rise in usage over the course of the past decade. Data scientist was dubbed the “sexiest job of the 21st century” by Harvard Business Review in October 2012 and many other publications have emphasised this new career type in recent years. Several individuals have taken an active interest in the field of data science in response to the significant rise in demand for data scientists and other professionals holding jobs with comparable titles.

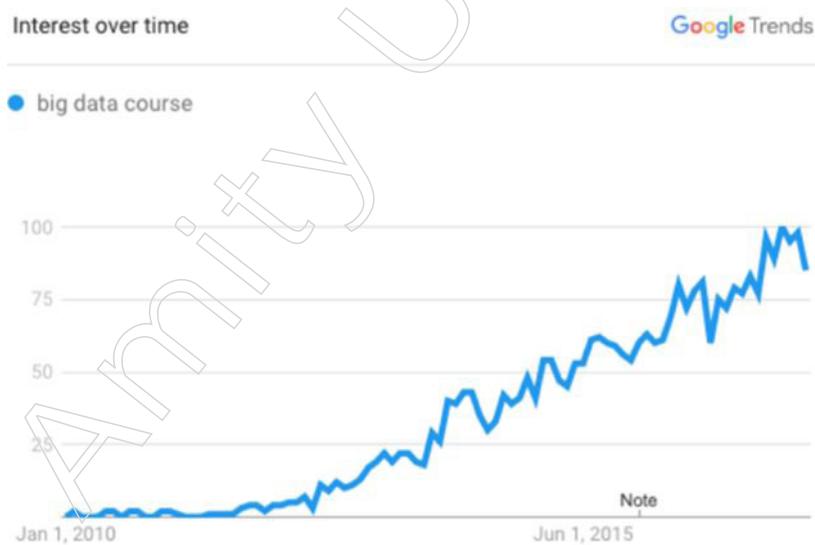


Figure 2: Increased knowledge about data science

As a result, the body of knowledge and education around data science has undergone a significant amount of professionalisation and each day, more information is made available to the public. The study of statistics and data analysis has traditionally

Notes

been confined to the realm of academia; but, in recent years, there has been a growing interest in the topic among both students and members of the working population.

Social Media Applications

The influence that social media has on people's lives is common knowledge at this point. On the other hand, the investigation of Big Data reveals that social media play a role of utmost significance. Not just because to the enormous number of data that is generated every day through social media platforms such as Twitter, Facebook, LinkedIn and Instagram, but also due to the fact that social media platforms give data on human behaviour in a virtually real-time format.

The data collected from social media platforms offer insights into the activities, tastes and opinions of "the public" on a scale that has never previously been known. Because of this, it is of tremendous value to anyone who is able to extract meaning from these massive amounts of data. Data collected from social media platforms may be put to a variety of uses, including the identification of consumer preferences for the purpose of product creation, the targeting of new customers for future purchases and even the targeting of potential votes in elections. It is possible that data collected from social media platforms may be regarded as one of the most important commercial drivers of big data.

The Upcoming Internet of Things (IoT)

The Internet of things (IoT) is a network of physical devices, vehicles, home appliances and other items that are embedded with electronics, software, sensors, actuators and network connectivity. This enables these objects to connect with one another and share data. Other terms for this network include the Internet of things (IoT) and the Internet of physical things (IoP). As manufacturers of consumer products begin incorporating 'smart' sensors into home appliances, its acceptance in the market is growing at an exponential rate. It was estimated that by the year 2020, the typical home will have a total of 50 internet-connected gadgets, as opposed to the about 10 devices that were present in the typical home in 2010. Thermostats, smoke alarms, TVs, audio systems and even smart refrigerators are a few examples of the types of gadgets that fall under this category.

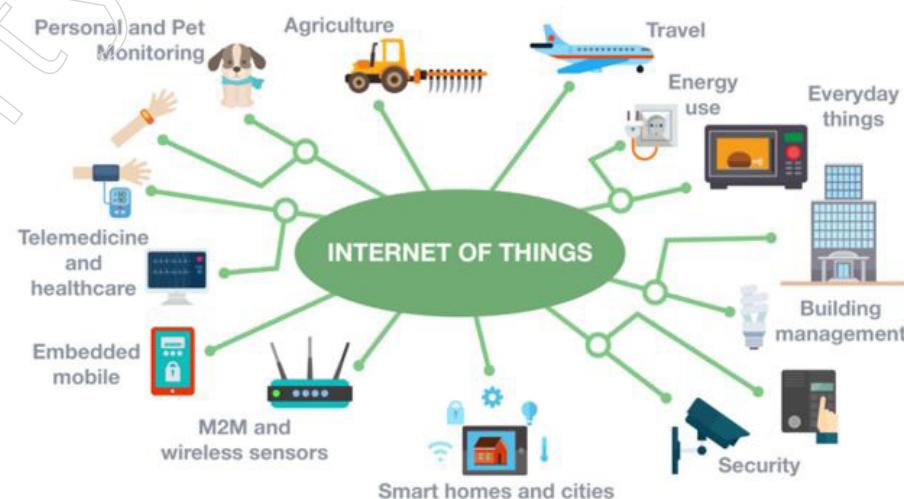


Figure 3: The Internet of things (IoT)

Each of these linked devices produces data, which is then sent between devices over the internet and which may be evaluated to obtain value. The data that is created through IoT devices, much like the data that is generated through social media platforms, is huge in terms of quantity and may give insights into the behaviour of customers. As a result of this, its value is exceptionally high.

1.2.4 Big Data and Data Science Hype; Datafication

- **Big Data and Data Science Hype**

Big Data is one of those overused and misunderstood buzzwords in the IT industry nowadays. It has developed into a catchphrase that may be used to describe any data-related issue. Little did we anticipate that the lack of clarity on whether it had to be all three of the Vs (volume, velocity and variety) offered the right fertile grounds for word abuse when the pundits defined the characteristics of Big Data and articulated the 3Vs (volume, velocity and variety). To add fuel to the fire, the absence of objective measures of what minimum Volume or Velocity would qualify as big data has led to the “beauty is in the eye of the beholder” syndrome, in which everyone comes up with their own qualifying criteria. This adds to the confusion that has already been caused by the lack of objective measures. In point of fact, there is a widespread tendency that when it comes to the first two Vs, Volume and Velocity, the criteria for justifying something as Big Data is almost anything that is more than what they are currently working with. This has become the standard for determining whether or not something can be considered Big Data. According to my understanding, this was on purpose (the lack of clear qualifying criteria).

It has been abundantly evident over the course of the years that the field of information technology, which is widely considered to be one of the most “self-sustaining” things that has ever materialised, has persistently built and pushed something new about every few years. Each had a significant chance to capitalise on the tens of billions of dollars that were being spent and invested to support the information technology sector. In order for any phenomena to have the necessary golden run, it is necessary for it to first go through a phase of hype, during which it must also maintain some degree of realism, until it reaches a critical mass of followers or adopters. Only if this resulted in enough excitement among executives about the possibilities that it began to impact their decision making would the golden run clock begin to count down. It starts with securing executive approval for allocating resources to experiment and investigate, which ultimately leads to significantly greater investments for larger scale deployments and initiatives.

It is quite evident that Big Data is now in its prime. This is something that is undoubtedly on the minds of a lot of company leaders all across the world right now. It is interesting to note that many non-tech savvy people do not even have a good understanding of what this is yet Big Data has garnered the status of a “competitive differentiator.” This is the pinnacle status for any phenomenon to achieve and only a select few such phenomena have ever achieved this. Big Data has garnered the status of a “competitive differentiator.” Due to the fact that big data can be applied in virtually every industry imaginable, it is an excellent contender for attaining this position.

Big Data may be linked with a certain amount of hype, but the applications and value that are swiftly becoming obvious lift the cloud of uncertainty and conjecture that

Notes

has been hanging over it. The era of big data, with its accompanying golden flight run, has already gotten off the ground. It is likely that, like past technologies of a similar nature, there will be numerous roller coaster moments, but overall, the journey is going to be a lot of fun. Don't let up (start working on one of these projects), fasten your seatbelt and enjoy the journey (you should and trust me, you will!). Buckle up, sit forward (identify the top possible candidates for big data projects) and don't relax (start working on one of these projects).

- **Concept of Datafication**

The process of transforming an organisation into a data-driven enterprise is referred to as "datafication," and it includes all of the tools, technology and procedures involved in this transformation. This term refers to an organisational trend that involves establishing the key to fundamental company operations through a worldwide reliance on data and the infrastructure that is associated with it. Datafication is sometimes referred to as datafy in some circles. One might say that a company or organisation is datafied if they have implemented datafication.

In order to successfully carry out essential business tasks, organisations need to collect data and extract knowledge and information. A company will also use data for making decisions, developing strategies and accomplishing other important goals. Datafication means that in today's world, which is more data-oriented, the existence of an organisation is dependent on having complete control over the storing, retrieving, manipulating and extracting of data and the information that is linked with it.

How to Datafy Your Business

The following are some strategies that can help you successfully datafy your business:

- a) **Use the Right Technology**

You may get started on datafication of your company by incorporating it into the Internet of Things (IoT). This necessitates having access to the appropriate technology, which may include mobile devices, wearables, Bluetooth beacons and voice assistants, amongst others.

- b) **Use the Appropriate Platform**

After you have the necessary infrastructure in place, the next step is to select the appropriate platform, which will determine how well you can extract data. This platform ought to generate the data that you want for the investigation. In addition to this, it should be able to transform vast volumes of data found online into information that is organised and usable by machines. If you choose the correct platform, it will provide you with the capabilities to monitor and analyse trends, which will in turn improve your ability to make decisions.

- c) **Build a Centralized Repository**

Only by utilising all of the data that is currently accessible will datafication be successful. As a result of this, you require a centralised repository to which all members of the company have access.

Current Applications of Datafication

Due to its many uses across a wide variety of sectors, like the ones listed below, "datafication" is no longer only a buzzword.

a) Human Resource Management

Mobile phones, social networking platforms and application data may all be mined by businesses for the purpose of locating new employees and doing in-depth analyses of their attributes, such as personality types and levels of comfort with taking risks. In place of requiring applicants to take personality tests, datafication can evaluate analytical thinking to determine whether or not individuals are a good fit for the corporate culture and the jobs for which they are applying. Datafication may result in the creation of new personality tests that may be utilised by human resources professionals.

b) Customer Relationship Management

Businesses that collect customer data may also gain a competitive advantage by utilising datafication technologies and techniques to better understand their consumer base. They are able to develop the necessary triggers that are relevant to the buying patterns and personalities of their target consumers. By the use of datafication, businesses are able to collect data based on the manner in which potential consumers communicate with the company via phone calls, emails and social media.

c) Commercial Real Estate

Those who work in the real estate sector, particularly those who specialise in commercial real estate, may also find value in datafication. Companies that deal in real estate can improve their understanding of a number of areas by making use of the tools and methods offered by datafication. As a result, they will be able to determine whether or not the parcel of land that they are considering is suitable for a customer who wants to launch a successful enterprise.

d) Financial Service Provision

It is possible that, of all the many types of businesses, the financial services sector stands to gain the most from datafication. Datafication is utilised by insurance companies in order to gain a better understanding of an individual's risk profile and to modernise their operational models. A person's capacity to repay a loan or mortgage may also be determined with this information, which is useful to the banking sector.

1.3 Statistical Inferences

Introduction

The act of analysing the results and drawing conclusions based on data that has been subjected to random fluctuation is known as statistical inference. Inferential statistics is another name for this method. The applications of statistical inference include the testing of hypotheses and the calculation of confidence intervals. The process of drawing conclusions about the characteristics of a population based on the results of a random sample is known as statistical inference. It is helpful in evaluating the link between the variables that are dependant and those that are independent. It is

Notes

the goal of statistical inference to arrive at an estimate of the uncertainty or the variance from sample to sample. Because of this, we are able to produce a probable range of values for the actual levels of anything that is prevalent in the population. The following factors are considered when drawing conclusions based on statistics:

- Sample Size
- Variability in the sample
- Size of the observed differences

Types of Statistical Inference

There are many distinct kinds of statistical inferences and many of them are utilised in the process of coming to conclusions. They are as follows:

- One sample hypothesis testing
- Confidence Interval
- Pearson Correlation
- Bi-variate regression
- Multi-variate regression
- Chi-square statistics and contingency table
- ANOVA or T-test

Statistical Inference Procedure

The following are the steps that are involved in inferential statistics:

- You should start with a theory.
- Formulate a working hypothesis for the research
- Ensure that the variables are operationalized.
- Identify the group of people to whom the findings of the study should be applicable.
- Provide a testable alternative to the null hypothesis for this group.
- Collect a representative sample of the population, then carry on with the research.
- Carry out statistical tests to determine whether or not the attributes of the gathered samples are sufficiently distinct from those that would be anticipated on the basis of the null hypothesis in order to be able to reject the null hypothesis.

Statistical Inference Solution

Statistical inference methods result in the effective utilisation of statistical data linked to populations of persons or experiments. It covers all of the characters, as well as the gathering, examination and analysis of data, as well as the organisation of the data that has been gathered. After beginning work in a variety of sectors, individuals are able to acquire information via the use of statistical inference solutions. The following are some examples of statistical inference solutions:

- Assuming that the observed sample is comprised of independent observations drawn from a Poisson or normal distribution population is a practise that is rather prevalent.

- The statistical inference solution is utilised in order to determine the value(s) of the parameter(s) of the anticipated model, such as the normal mean or the binomial proportion.

Importance of Statistical Inference

Inferential statistics are necessary for conducting an accurate analysis of the data. Interpreting the findings of the research requires careful data analysis to ensure an appropriate conclusion can be drawn from it. Its primary use is in the forecasting of future events for a wide range of data in a variety of domains. It facilitates the process of drawing conclusions based on the facts. The statistical inference can be used in a variety of contexts, including but not limited to the following areas:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector

Common Inferential Methods

The following are four common practises that may be used to draw conclusions based on statistical data.

- Hypothesis Testing: This method use representative samples to evaluate two hypotheses about a population that are incompatible with one another. After taking into consideration the possibility of sampling error, results that are statistically significant imply that the sample effect or link does exist in the population.
- Confidence intervals are defined as a set of values that most likely contains the population value. At this step, the sample error is assessed and a margin is added around the estimate. This provides a notion of how incorrect the estimate may potentially be.
- Margin of Error: This is quite similar to a confidence interval, however it is more commonly used for survey findings.
- Regression Modelling: Estimation of the process in the population that is responsible for producing the results, as seen in regression modelling.

1.3.1 Role of Statistics in Data Science, Inference Types

How is Statistics Currently Used in Data Science?

The fields of data science and data analytics continue to rely heavily on statistical methodology. The following is a small selection of the several ways in which this field of mathematics is assisting data analysts in their work with large amounts of data:

- **Testing hypotheses:** Putting hypotheses to the test is an essential part of the analytics process that should not be overlooked. The purpose of a hypothesis

Notes

test is to determine whether or not two claims about a certain population that are mutually exclusive are true. This test is a useful instrument for determining whether or not a discovery is significant from a statistical point of view.

- **Creating probability distributions and estimation:** When statistical approaches are used to data, this helps in the creation of probability distributions and estimations, which in turn may assist generate a better understanding of logistic regressions and machine learning.
- **Informing business intelligence:** Statistics are frequently utilised for a variety of business activities because they give a degree of confidence in the results, which can then be used for making forecasts and projections.
- **Creating learning algorithms:** Algorithms such as naive Bayes and logistic regression have progressed throughout time to accommodate the requirements of data analysis.
- **Aiding with prediction and classification:** Assisting with forecasting and categorising data Statistics is a strong instrument that may be utilised for forecasting and categorising data.
- **Incorporating descriptive statistics:** The use of descriptive statistics provides descriptions and summaries of data, in addition to visualisation options that allow the insights to be presented to a non-technical audience in an easy-to-understand manner. Incorporating descriptive statistics is one of the most important aspects of data analysis.
- **Determining probability:** Statistical formulae relevant to probability have a wide variety of applications. One of these is determining probability. Some examples of this include clinical studies, political polls, actuarial tables and even the calculation of the probability that a population would get infected with a disease.

Application of Statistics in Data Science:

The following is a list of key ideas that every Data Scientist and Analyst ought to be familiar with in order to do their jobs effectively:

1. Classification

The umbrella word for the processes involved in data mining is classification. During this stage of the process, we sort the data into different subsets according to a variety of criteria. These characteristics may be ones that we discovered via study; they could be dependent on our aims; and finally, we could sectionalize the data utilising patterns identified in data visualisation and sampling. The process of classification, which is also known as a decision tree, may be accomplished using three important approaches: linear discriminant analysis, logistic regression and k-nearest neighbours.

Classification is a common application in Data Science. Data Scientists and Analysts are always tasked with finding new methods to categorise emails as “spam” or “important.” Similarly, AI will categorise news articles based on your prior searches as well as the amount of time you spend reading each article, among other considerations. Nevertheless, the strategies for categorization are not restricted to the three that have been explained above. You would need to consistently improve both your system and

your methodology in order to make the most accurate predictions possible on the qualitative replies. If you are good at programming and are interested in working in the field of data science, taking statistics classes online may be quite helpful and save you a lot of time.

2. Logistic Regression

Resampling is a common technique used for conducting objective and accurate analyses of huge data sets. During the examination of enormous volumes of data, the method removes the possibility of error associated with the parameters of the population.

This approach repeatedly selects samples from a large amount of data in order to generate a specific and one-of-a-kind sampling distribution that accurately reflects the data in question. This method takes into account all of the conceivable outcomes of the investigation, which enhances accuracy while also reducing bias. This approach repeatedly selects samples from a large amount of data in order to generate a specific and one-of-a-kind sampling distribution that accurately reflects the data in question. This method takes into account all of the conceivable outcomes of the investigation, which enhances accuracy while also reducing bias.

3. Resampling Methods

Resampling is a standard method to analyse large data samples unbiased and precise. The technique eliminates the uncertainty of population parameters during the analysis of massive amounts of data.

The method continually draws out samples from extensive data to obtain a small and unique sampling distribution that represents the original data. The technique covers all possible results of research and thus improves accuracy and decreases bias.

The method continually draws out samples from extensive data to obtain a small and unique sampling distribution that represents the original data. The technique covers all possible results of research and thus improves accuracy and decreases bias.

Advantages of Learning Statistics for Data Science

1. Helps organise the data

While developing their marketing strategies, businesses really need to ensure that the data they collect has been correctly categorised. Not only that, but the organisation of data into categories and hierarchies also assists businesses in making targeted improvements to their goods and services. In data science, having data that is not structured makes it impossible to use, which is a loss of both time and an asset.

2. Helps Spot Trends

The process of data collecting may be hard on all three of your mental, physical and financial resources. You may save a significant amount of time and money by conducting targeted research. The early identification of trends through the use of statistics provides data scientists with the ability to more precisely target their areas of investigation.

Notes

3. Helps in Estimation and Probability Distribution

The understanding of logistic regression, cross-validation and other such algorithms are the foundation of both data analytics and machine learning. These methods assist the machine in predicting the next move you will take. When you think about the recommendations that appear when you are listening to music on YouTube, you will realise that there are at least a few songs that you would appreciate even if you have never heard them before.

4. Makes Data Visualization Easier

When it comes to big data analysis, visualisation techniques such as histograms, pie charts and bar graphs go a long way right to the top to make data more interactive and meaningful. They render the understanding of complicated data in a manner that is both interactive and simple to grasp. These statistical methods assist in the early detection of patterns and make them understandable to even the untrained eye. As a direct consequence of this, drawing conclusions and formulating action plans become less difficult.

5. Help Reduce Assumptions

Knowledge of mathematical analysis, namely differentiation and continuity, is necessary for understanding the fundamentals of artificial intelligence (AI), machine learning and data analytics. These elements contribute to more accurate predictions of outcomes, which are based on inferences rather than assumptions. Statistics reduces the number of assumptions, which ultimately results in an improvement in the model's capacity for prediction. We haven't arrived at this point when a significant portion of what we see is pertinent and likely connected to what we want to see by some sort of mystical force.

6. Help Account for Variability in Data

In model-based data analytics, statistics may be used to take into consideration a variety of factors, such as clusters, time, geography, etc. If statistical procedures are not utilised, then an analysis of the data may take place without taking into consideration the variability of the data, which may lead to the production of inaccurate estimations.

When you have a good understanding of the techniques of distribution, you also have a better understanding of the variable components. In addition to visualisation, one of the most important aspects of data analytics and statistics is, as is only natural, the mechanism by which the data is presented.

1.3.2 Population and Samples, Statistical Modelling

- Population and Samples

In statistics, as well as in quantitative methodology, a collection of data is gathered and selected from a statistical population with the assistance of some established processes. This process takes place both in statistics and in quantitative methodology. There are two distinct varieties of data sets, which are referred to respectively as population and sample. When we calculate the mean deviation, variance and standard deviation, it is necessary for us to know whether we are referring to the entire population or just to the sample data. If we are only referring to the sample data, then

the mean deviation, variance and standard deviation will be incorrect. If the size of the population is expressed by the letter 'n,' then the size of the sample taken from that population is represented by the number $n - 1$. Let us take a closer look at the data sets for the population as well as the data sets for the samples.

Population

It contains all the components from the data set and properties of the population that can be measured, such as the mean and the standard deviation, are referred to as parameters. As an illustration, the phrase "All people living in India" refers to the whole population of India.

There are several subgroups that make up the total population. They are as follows:

- Finite Population
- Infinite Population
- Existential Population
- Hypothetical Population

Let us go through each of the categories one at a time.

a) Finite Population

Another name for a population that can be counted is a countable population and the limited population falls under this category. To put it another way, it is the population of all the persons or objects that have a limited number of occurrences. When doing statistical analysis, having a population that is either finite or infinite, rather than both, is preferable. Employees of a firm and potential customers in a market are two examples of populations that are examples of finite populations.

b) Infinite Population

Another name for an endless population is an uncountable population, which refers to a population in which it is impossible to count the individual units that make up the population. An unlimited number of germs is one illustration of an example of an infinite population that can be found in a patient's body.

c) Existential Population

The population of actual living people is what demographers mean when they talk about the "existing population." In other words, the population for which there is a unit that can be obtained in a tangible form is referred to as the "existing population." Books, students and other things are examples.

d) Hypothetical Population

The term "hypothetical population" refers to a population for whom the "unit" in question is not readily available in a tangible form. A population is made up of groups of observations, items, or other things that share some characteristic in common. In certain circumstances, the populations can only be imaginary. Examples of random outcomes include the results of rolling dice or flipping a coin, for instance.

Notes

Sample

It consists of one or more observations that are drawn from the population and the attribute that may be measured about a sample is referred to as a statistic. The process of choosing a representative sample from a population is referred to as sampling. For instance, the sample of the population would consist of some persons now residing in India.

There are essentially two different forms of sampling. They are as follows:

- Probability sampling
- Non-probability sampling

Probability Sampling

In probability sampling, the units of the population being sampled are not chosen at random by the researcher like in other sampling methods. This may be handled by following particular processes, which will guarantee that each unit of the population has a set probability of being included in the sample. This will ensure that accurate results can be obtained. Taking samples in this manner is sometimes referred to as random sampling. The following are examples of some of the methods that may be used for conducting probability sampling:

- Simple random sampling
- Cluster sampling
- Stratified Sampling
- Disproportionate sampling
- Proportionate sampling
- Optimum allocation stratified sampling
- Multi-stage sampling

Non-Probability Sampling

The researcher has complete freedom over which members of the population to sample using the non-probability sampling method. The selection of units for these samples will be based only on human opinion and there is no theoretical foundation upon which to base an estimate of the population's characteristics. Non-probability sampling can be carried out using a variety of methods, such as:

- Quota sampling
- Judgement sampling
- Purposive sampling

Characteristics of the sample

In order to be suitable for statistical analysis, a sample need to exhibit a set of defining criteria. If research is conducted on an inaccurate sample, the resultant conclusions will be incorrect and these may lead to severe repercussions since they would contradict the behaviour of the entire community.

1. Representativeness

The behaviour of a population as a whole ought to be reflected accurately in a sample. Consider the circumstance described in the previous example, in which 5,000 employees out of 50,000 employees are chosen for the position. If there are 40,000 male employees in the population as a whole, but only 40,000 female employees were included in the sample, what would be the results? (which is the sample size). Any study that is done based on this sample will not accurately represent the behaviour of the population as a whole.

2. Homogeneity

The consistency of behaviour over a number of different samples is what we mean when we talk about homogeneity. If we take many samples from the same population, we should anticipate that those samples will all arrive at similar inferences about the population as a whole.

Assume that we have three samples, each of which has a sample size of 5,000 and that we wish to determine the mean wage of the 50,000 employees.

- The average annual income for Sample 1 is \$40,000
- The average annual income for Sample 2 is \$38,000.
- The sample with the mean salary of \$41,000 is sample number 3.

We are able to classify these samples into the category of being homogenous since all of the samples provide roughly equivalent information with regard to the salaries of the workers.

What if the outcome turns out to be anything like this,

- The average annual income for Sample 1 is \$40,000
- The sample with the mean salary of \$15,000 is sample number 2.
- Sample 3 has a mean salary of one hundred thousand dollars.

Due to the instability of the data, the researcher won't be able to calculate the approximate wage of a person working for the firm in this case.

3. Adequacy

It is important that the number of sampling units in a sample be sufficient so that the study may be carried out. In the previous illustration, if we were to do research with a sample size of five or six, it would not be efficient because there are fifty thousand employees in total.

4. Similar Regulating Conditions

If many samples are required, there need to be an analogous procedure for picking them out. In the previous illustration, a sample of 5,000 employees was picked at random from a total of 50,000 employees; similarly, if we are picking another sample, it ought to be chosen at random as well. There should be no encouragement of any form of pre-conditions for the selection of the elementary unit. In the event that Sample 1 of the sample size 5k is selected at random, but Sample 2 of the same sample size

Notes

is created for the same data analysis, with the exception that Sample 2 is comprised of solely female employees, what would the results be? This will have an impact on the homogeneity of the samples, which will lead to inaccurate conclusions being drawn from the data.

- **Statistical Modelling**

Statistical modelling is an involved process that involves producing sample data and making predictions about the actual world by employing a large number of statistical models and making explicit assumptions. There is a mathematical connection that can be made between the random and non-random variables that are involved in this process. Data scientists are able to identify the correlations between random variables and analyse information in a strategic manner because to this capability.

When statistical models are applied to raw data, intelligible visualisations may be produced using statistical models. These visualisations allow data scientists to uncover relationships between variables and provide predictions. For the purposes of statistical analysis, some examples of common data sets are census data, statistics on public health and data from social media.

Statistical Modeling Techniques in Data Analysis

The collecting of data is the first step in the statistical modelling process. The information might have come from the cloud, spreadsheets, databases or some other sources. The statistical modelling approaches that are employed in the study of data may be divided into two groups. These include:

Supervised Learning

In the case of the supervised learning model, the algorithm learns from a data set that has been labelled. This data set also includes an answer key, which the programme uses to assess how accurately it is learning from the data. Techniques of supervised learning used in statistical modelling include the following:

- **Regression Model:** A prediction model that is meant to analyse the connection between an independent variable and a dependent variable is called a regression model. The logistical, polynomial, and linear regression models are the ones that are used most frequently. The link between variables, modelling, and forecasting may be determined with the use of these models.
- **Classification Model:** A complicated and extensive collection of data points are analysed and categorised using a classification model that is driven by an algorithm. The decision tree, the Naive Bayes model, the closest neighbour model, the random forest model, and the neural networking model are all examples of common models.

Unsupervised Learning

The unsupervised learning model provides the algorithm with data that has not been labelled, and the system then makes independent efforts to extract features and discover patterns from the data. Unsupervised learning is demonstrated through things like clustering algorithms and association rules. Here are two illustrations of this:

- **K-means clustering:** It is a type of data analysis that uses an algorithm to organise a certain number of data points into distinct categories on the basis of their similarities.
- **Reinforcement learning:** It is a method that includes training an algorithm to iterate over numerous trials using deep learning, rewarding actions that result in favourable results, and penalising activities that create unwanted consequences.

How to Build Statistical Models

One of the most difficult aspects of teaching statistics is model development, which entails selecting appropriate predictors. It is tough to write down the processes because at each stage, you must analyse the situation and decide what action to do next. This makes it difficult to write down the processes. If you are only interested in running predictive models and don't care about the links between the variables, this will be a lot simpler for you to execute. Go forward with the model of regression using steps. Let the data determine what the most accurate forecast is for you. If, on the other hand, the objective is to provide a solution to a research topic regarding relationships, you will need to get your hands filthy.

Step 1

The first thing you'll need to do is choose the statistical model that caters to your requirements the most effectively. To begin, you will need to determine whether you will respond to a single enquiry or provide a forecast based on a huge number of different parameters. Take into account the number of independent and dependent variables that can be accessed. What is the minimum number of variables that must be incorporated into the model? What is the link between the variables that are being explained and the variables that are being explained by?

Step 2

As soon as you've settled on a statistical model, you should go right into the descriptive statistics and visualisations. Creating a visual representation of the data can make it easier for you to spot errors and have a better knowledge of the variables and the behaviour they exhibit. Construct predictors to investigate the ways in which related variables interact with one another and the outcomes of combining datasets.

Step 3

It is essential that you have a solid understanding of the connection that exists between the potential predictors and the correlation that they have with the findings. For this, you need to keep an accurate record of the results, whether or not there were control variables involved. You might also remove variables from the model that are not significant in the beginning while keeping all of the other variables.

Step 4

Throughout the process of analysing the current correlations between variables, as well as evaluating and categorising every potential predictor, you can keep the essential research questions in mind.

Notes

Step 5

With statistical modelling software, one may collect data, organise that data, analyse that data, interpret that data and build new analyses. This software package includes data visualisation, modelling and mining capabilities, all of which contribute to the overall level of process automation.

1.3.3 Probability Distribution: Types and Role

The multidisciplinary discipline of data science has recently seen a rise in its appeal. It does this by employing scientific methodologies, methods, algorithms and tools in order to extract facts and insights from organised, semi-structured and unstructured databases. These data and insights are used by businesses to enhance productivity, extend their operations and better predict the demands of their customers. While undertaking data analysis and constructing a dataset for use in model training, it is vital to pay attention to the probability distribution.

What Is Probability?

The term “probability” refers to the likelihood that something will take place. It is a mathematical concept that estimates the likelihood that certain occurrences will take place. The probability values are presented between 0 and 1, with 0 being the most likely outcome. The degree to which an occurrence is likely to take place is what is meant when we talk about probability. This fundamental theory of probability may also be applied to probability distributions in a variety of contexts.

What Are Probability Distributions?

A probability distribution is a statistical function that specifies all of the potential values and probabilities for a random variable within a specific range. This function is known as a probability distribution. This range will be bounded by the smallest possible value and the greatest possible value; however, the location on the probability distribution where the potential value would be displayed will be governed by a number of other factors. These characteristics include the mean (or average), standard deviation, skewness and kurtosis of the distribution.

Types of Probability Distribution

The probability distribution is divided into two parts:

1. Discrete Probability Distributions
2. Continuous Probability Distributions

Cumulative Probability Distribution

There is another name for the cumulative probability distribution and that name is the continuous probability distribution. The collection of outcomes that are feasible under this distribution can take on values that are anywhere along a continuous scale.

One example of a continuous or normal distribution is a collection of real numbers since these numbers can take on any one of an infinite number of conceivable forms. In the same vein, some instances of the Normal Probability distribution include a collection

of complex numbers, a collection of prime numbers, a collection of whole numbers, etc. In addition, one example of continuous probability that may be found in real-life situations is the temperature of the day. A distribution table may be constructed with these findings as the basis. It may be described using a probability density function. To calculate the normal distribution, the formula is as follows:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where,

- μ = Mean Value
- σ = Standard Distribution of probability.
- If $\text{mean}(\mu) = 0$ and $\text{standard deviation}(\sigma) = 1$, then this distribution is known to be normal distribution.
- x = Normal random variable

Normal Distribution Examples

Statistics based on the normal distribution have become the de facto standard for many different kinds of probability questions due to the accuracy with which they estimate a variety of natural occurrences. The following are some examples:

- Height of the Population of the world
- Rolling a dice (once or multiple times)
- To judge the Intelligent Quotient Level of children in this competitive world
- Tossing a coin
- Income distribution in countries economy among poor and rich
- The sizes of females shoes
- Weight of newly born babies range
- Average report of Students based on their performance

Discrete Probability Distribution

If the possible outcomes can be broken down into separate parts, then the distribution in question is known as a discrete probability distribution. If one rolls a dice, for instance, all of the potential results are discrete and add up to a large number of possibilities. This creates a mass of outcomes. The probability mass function is another name for this concept. So, the results of a binomial distribution are made up of n separate trials in which the outcome may or may not take place. In the case of the binomial distribution, the formula is as follows:

$$P(x) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

$$P(x) = C(n, r) \cdot p^r (1-p)^{n-r}$$

Where,

- n = Total number of events
- r = Total number of successful events.

Notes

- p = Success on a single trial probability.
- ${}^nC_r = [n! / r!(n-r)!]$
- $1 - p$ = Failure Probability

Binomial Distribution Examples

As is common knowledge, the binomial distribution allows for a variety of outcomes to be possible. In actual practise, the principle is applied in the following contexts:

- Throughout the process of producing a product, it is necessary to determine the total amount of materials that were either utilised or discarded.
- To conduct a poll asking individuals about the good and negative feedback they have on something.
- To determine whether or not a specific channel is seen by a certain number of people by computing the results of a poll asking "Yes" or "No."
- The proportion of male to female employees in a certain organisation.
- The process of tallying the votes cast for each candidate in an election, among other similar tasks.

What are the 3 Probability Distributions Used in Data Science?

In the field of data science, probability distributions are an essential subject since they are used to better comprehend the patterns that may be found in data. The following are some examples of three types of probability distributions that are utilised often in data science:

1. Normal Distribution:

The normal distribution, which is also known by its other name, the Gaussian distribution, is a continuous probability distribution that is widely used to model occurrences that take place in the actual world. When we move further out from the centre of the distribution, the frequency of the data points drops and the distribution takes on a bell-shaped and symmetrical structure.

The normal distribution is utilised in a variety of statistical procedures, such as testing hypotheses and creating confidence intervals. It is also beneficial for generating predictions about the future based on data from the past since it provides a reliable estimate of the data's central tendency.

This probability distribution is symmetrical with respect to its mean value. In addition to this, it shows that data that is close to the mean happens more frequently than data that is relatively distant from it. Here, the mean is 0, and the variance is a finite value.

In the example, you produced one hundred random variables with values between one and fifty. After that, you developed a function in order to construct the normal distribution formula in order to calculate the probability density function. The data points and the probability density function are then shown against the X-axis and the Y-axis, respectively.

Notes

```

import numpy as np
import matplotlib.pyplot as plt

# Creating a series of data of in range of 1-50.
x = np.linspace(1,50,100)

#Creating a Function.
def normal_dist(x , mean , sd):
    prob_density = (np.pi*sd) * np.exp(-0.5*((x-mean)/sd)**2)
    return prob_density

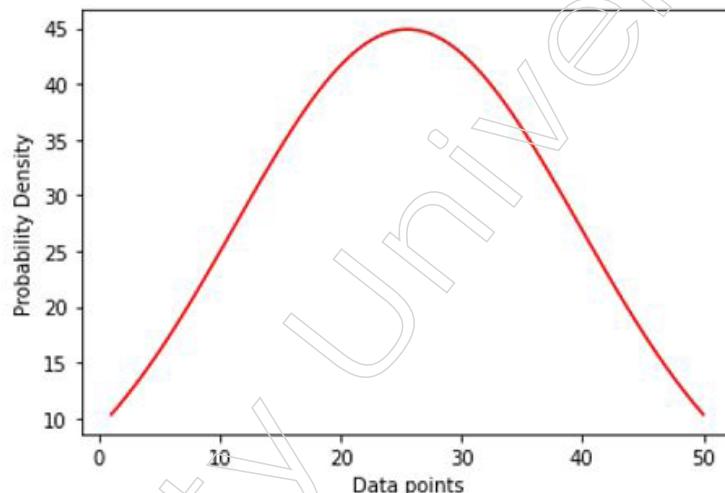
#Calculate mean and Standard deviation.
mean = np.mean(x)
sd = np.std(x)

#Apply function to the data.
pdf = normal_dist(x,mean,sd)

#Plotting the Results
plt.plot(x,pdf , color = 'red')
plt.xlabel('Data points')
plt.ylabel('Probability Density')

Text(0, 0.5, 'Probability Density')

```

**2. Poisson Distribution:**

The Poisson distribution is a discrete probability distribution that is used to mimic the frequency of occurrences of an event over a specific length of time or space. This simulation may be done in a variety of different ways. It has widespread use in fields such as biology, physics and economics, all of which deal with the occurrence of random and unrelated occurrences. The mean and standard deviation of the Poisson distribution are identically calculated values.

As a result, the probability of seeing an event is highest when we are closest to the mean and it decreases as we go further away from the mean. Estimating the chance of rare events, such as breakdowns in manufacturing processes or the number of customers that walk through a company's doors, may be done with the help of the Poisson distribution.

Notes

A straightforward illustration of the Poisson distribution may be seen in the following Python code.

It has two parameters, which are:

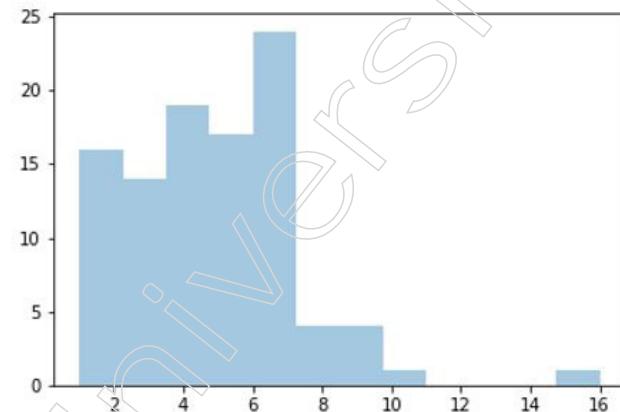
1. Lam: Known number of occurrences
2. Size: The shape of the returned array

The 1x100 distribution for occurrence 5 will be generated using the Python code that is given below.

```
from numpy import random
import matplotlib.pyplot as plt
import seaborn as sns

sns.distplot(random.poisson(lam=5, size=100), kde=False)

plt.show()
```



3. Binomial Distribution:

The discrete probability distribution known as the binomial distribution is used to estimate the proportion of endeavours that are fruitful given a predetermined total number of tests. It is used rather frequently in sectors such as marketing, quality control and medical, all of which provide binary results, as an example.

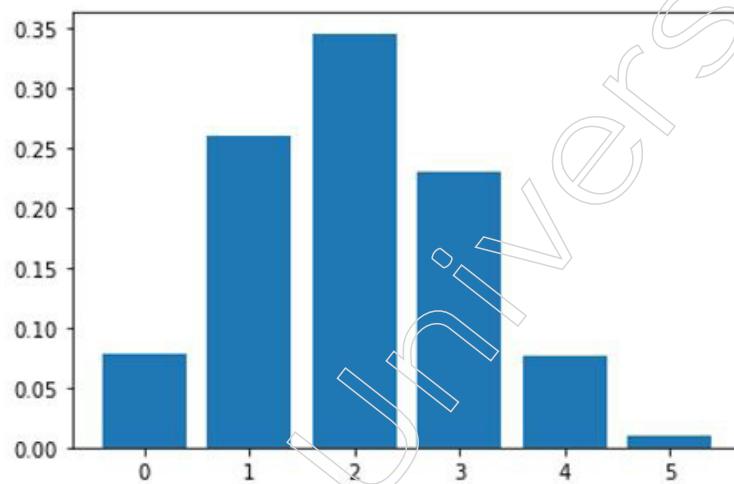
The binomial distribution is characterised by two parameters: the first is the number of trials involved and the second is the probability that each trial will be successful. It is helpful for calculating the likelihood of a particular number of successes in a certain number of trials and it may be put to use to test hypotheses concerning the proportion of successful occurrences within a population.

The binomial distribution is a discrete distribution, meaning that there are only a set number of different outcomes possible. The binomial distribution is revealed by viewing a sequence of Bernoulli trials, which are also known as Bernoulli experiments. The results of a scientific experiment known as a Bernoulli trial can only be one of two things: successful or unsuccessful.

Take for example a randomised experiment in which you have a 0.4 percent probability of getting heads when you toss a biased coin six times. If "getting a head"

is taken into account when determining what constitutes a “success,” then the binomial distribution will display the probability of r successes for each possible value of r . The number of successful Bernoulli trials (r) in a series of n consecutively independent trials is represented by the binomial random variable.

```
from scipy.stats import binom
import matplotlib.pyplot as plt
# setting the values
# of n and p
n = 5
p = 0.4
# defining list of r values
r_values = list(range(n + 1))
# list of pmf values
dist = [binom.pmf(r, n, p) for r in r_values]
# plotting the graph
plt.bar(r_values, dist)
plt.show()
```



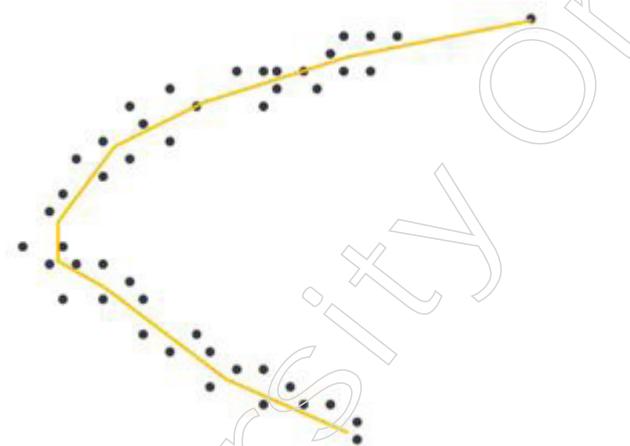
1.3.4 Fitting a Model

Model fitting is a measurement of how effectively a machine learning model adapts to data that is comparable to the data on which it was trained. This allows the model to perform as well as it did when it was trained on the original data. In most versions, the procedure for fitting is pre-programmed and takes place automatically. When fresh data are input into a model that is well-fit, the model will produce results that are more accurate by providing an accurate approximation of the output. Fitting a model involves modifying the parameters contained inside the model, which ultimately results in increased precision. During the process of fitting, the algorithm is executed on the test data, which is sometimes referred to as the “labelled” data. In order to determine whether or not the model is accurate, the outputs of the algorithm need to be compared to real and observed values of the target variable once the algorithm has completed its execution. Utilizing the findings, more adjustments may be made to the algorithm’s parameters in order to improve the uncovering of links and patterns between the inputs, outputs and targets. The procedure might be carried out several times up until the point when reliable and precise insights are obtained.

Notes

What Does a Well-Fit Model Look Like?

A model that is well-fitted should not only closely match the data that is currently available, but it should also closely follow the overall contours of the model. No model will ever be able to match the input data exactly, but a well-fit model will be able to match the data and the basic shapes in a way that is quite similar. Note that the line in the figure below does not perfectly match each individual data point, but it does reflect the broad curve. This is an essential aspect to keep in mind.



Why is Model Fitting Important?

As was said before, a model that is accurate does not match each and every data point that is provided, but rather it follows the general trends. It may be concluded from this that the model is not underfit nor overfit. If a model is not well fitted, it will generate inaccurate insights and you should not utilise it for making judgements because of this.

Underfitting

Underfitting is when a model oversimplifies the data and fails to capture sufficient information about the relationships that are there within the data. This is typically the result of not enough time being spent training the models. When a model has a poor performance on the data used for training it, this is an indication that the model is underfit.

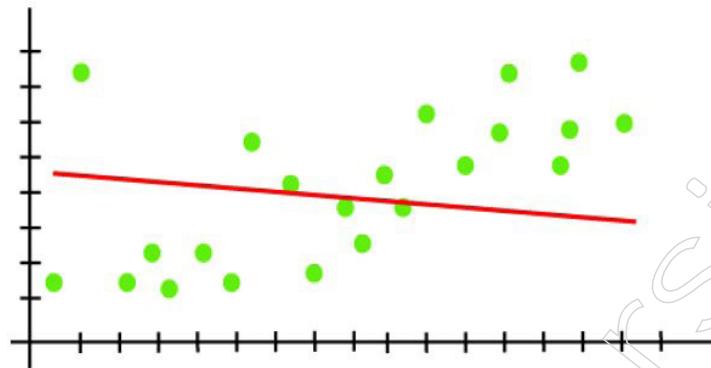
If it does occur, then it simply indicates that our model or method does not provide a sufficient enough fit to the data. It occurs most frequently when we have a limited amount of data upon which to construct an appropriate model, as well as when we attempt to construct a linear model with a limited number of non-linear data. In situations like these, the rules of the machine learning model are overly simple and versatile to be applied to such a little amount of data, and as a result, the model will most likely provide a large number of inaccurate predictions.

In a word, underfitting describes a situation in which a model neither generalises effectively to new data nor performs well on the data that it was trained on.

Reasons for Underfitting

- Having a high bias while having a low variance
- The size of the dataset that was utilised for training was insufficient.

- The model is far too straightforward.
- The training data have not been cleaned, and they also have noise in them.
- Techniques to reduce underfitting:
- Increasing the model's degree of complexity
- Feature engineering should be performed when the number of features is increased.
- Reduce the amount of noise in the data.
- If you want better outcomes, either increase the number of epochs you train for or the total amount of time you train for.



Overfitting

Contrast this with overfitting, which is the reverse of underfitting. It occurs when a model is extremely sensitive to the data that it contains, which leads to an over-analysis of the patterns that are included inside the model. In most cases, overfitting occurs as a direct consequence of excessive practise on training data sets. It may be recognised when a model performs well on the data that was used for training, but when it is presented with new data, it works badly and does not adapt to it.

Non-parametric and non-linear approaches are responsible for the phenomenon of overfitting. This is due to the fact that these types of machine learning algorithms have more flexibility in the process of generating the model based on the dataset, and as a result, they are more likely to produce models that are unrealistic. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

Overfitting is an issue that arises when the assessment of machine learning algorithms on training data is different from the evaluation on data that has not been seen before.

Reasons for Overfitting are as follows:

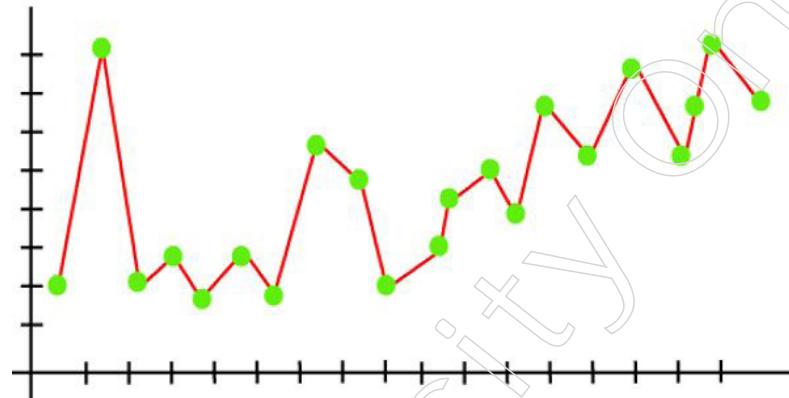
- High levels of variation and low levels of bias
- The model is far too complex.
- The quantity of the data used for training

Techniques to reduce overfitting:

- Obtain additional data through training.

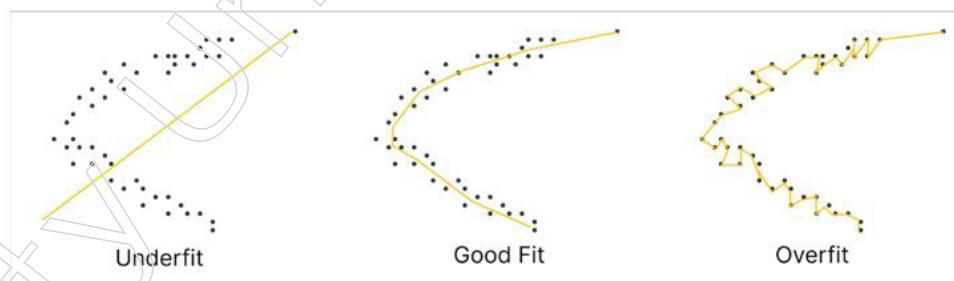
Notes

- Reduce the complexity of the model.
- Early termination of the training phase (keep a close check on the loss during the training time; as soon as it begins to grow, terminate the training immediately).
- Ridge Regularisation and Lasso Regularisation.
- Use dropout for neural networks to tackle overfitting.



A Well-Fit Model

A model that is well-fit has accurate hyperparameters that represent the relationships between the variables and the target variables and as a result, it performs well both on the data used for training and on the data used for evaluation. In most cases, fitting is an automated process that involves the hyperparameters being modified individually and automatically so that they are the best possible match for the data that is supplied. Users are able to improve their decision-making and obtain more accurate insights when they utilise models that are a good fit.



1.4 Introduction to R and Information Visualisation

Introduction

- **R**

R is a programming language and environment that may be used for statistical computing and graphical representation. It is a project that is being developed by GNU that is analogous to the S programming language and environment that was created at Bell Labs (which was then AT&T and is now Lucent Technologies) by John Chambers and his colleagues. One way that S may be implemented is through R, but there are other ways as well. While there are some significant changes, the majority of the code that was created for S may be run without modification in R.

R is a highly extensible programming language that offers a wide range of statistical (including linear and nonlinear modelling, traditional statistical tests, time-series analysis, classification and clustering) and graphical (including those methods). When conducting investigations into statistical methods, the S programming language is frequently the instrument of choice and R offers an Open Source entry point for taking part in such investigations.

One of the benefits of using R is how straightforward it is to generate charts of publication-level quality, complete with appropriate mathematical symbols and formulas. This is one of R's many capabilities. The default settings for the most insignificant design decisions in visuals have been given a lot of attention, but the user still has complete control.

The source code for R may be downloaded as free software and is distributed under the GNU General Public License, which is governed by the Free Software Foundation. It may be compiled and executed on a broad range of UNIX platforms and other related operating systems, such as Windows and Macintosh, as well as FreeBSD and Linux.

- **Information visualization**

The process of portraying data in a meaningful and visually appealing fashion that end users can readily perceive and grasp is referred to as "information visualisation." This includes graphical representations of data as well as dashboards. The successful communication of insights to non-experts in a format that is easily consumable may be accomplished using information visualisation. In most cases, it demonstrates pertinent linkages in the data, which enables decision-makers to draw inferences and behave in a manner that is informed by the information more readily.

1.4.1 R Windows Environment, its Data Type, Functions, Loops, Data Structure

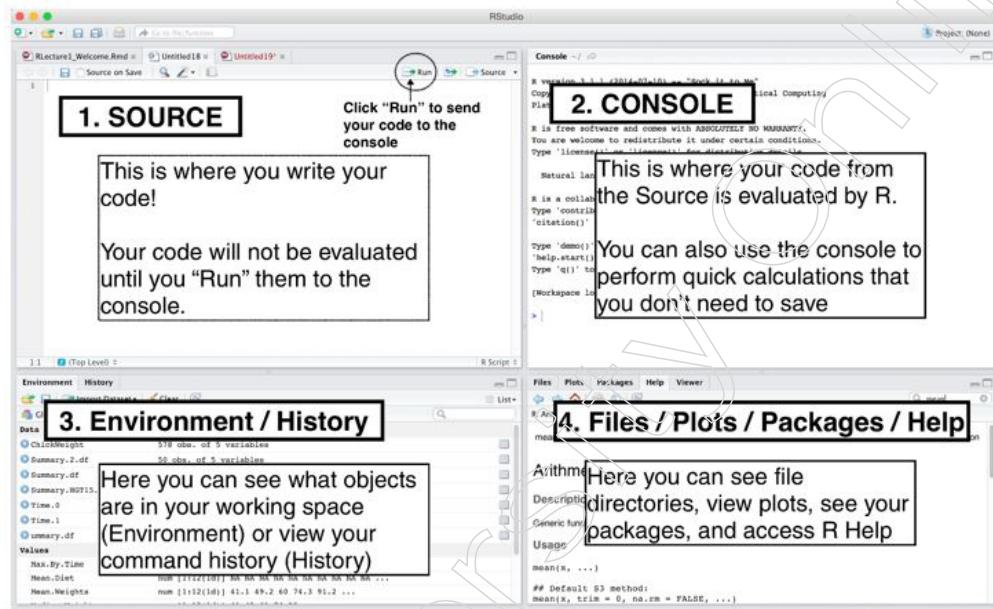
R is a freely available programming language that has found widespread application as statistical software and a data analysis tool. The Command-line interface is often included with R installations. R is a programming language that can run on a variety of operating systems, including Windows, Linux and macOS. In addition, the R programming language is the most advanced and cutting-edge technology currently available.

Ross Ihaka and Robert Gentleman from the University of Auckland in New Zealand were the ones responsible for its creation, while the R Development Core Team is the one responsible for its ongoing development. An implementation of the R programming language may also be found in the S programming language. In addition to this, it incorporates lexical scoping semantics that are derived from Scheme. In addition, the idea for the project was conceived in 1992, the first version was published in 1995 and the beta version was made stable in the year 2000.

When an interpreter for a programming language is started up, a virtual area known as the environment is created and made available for use. To put it another way, the environment is just a collection of all the functions, variables and objects. Instead, Environment may be thought of as a top-level object that stores the collection of names/variables that are connected to a certain set of values. In this post, we will

Notes

talk about how to create a new environment in R programming, as well as how to list all environments, how to remove a variable from an environment, how to search for a variable or function among environments and how to search for function environments using examples.



Source: It enables R to take its input directly from the specified file, URL, connection, or expressions in the program's environment. The input is read from that file and then processed until the end of the file is reached; after that, the parsed expressions are evaluated in the environment of your choice in the order that they were read.

Console: R will display the results of a command in the console window, which is located in the bottom left panel of RStudio. This is the location where R is waiting for you to tell it what to do and where you may enter commands. You are able to input instructions straight into the console; however, such commands will be lost when the session is terminated.

Environment: The environment is a virtual space that is triggered when an interpreter of a programming language is launched. To put it another way, the environment is just a collection of all the functions, variables, and objects. Alternately, Environment may be thought of as a top-level object that stores the collection of names/variables that are connected to a certain set of values.

Files: You may gain access to the file directory on your hard drive by using the Files panel. You may set your working directory using the "Files" panel, which is a useful function. Once you have navigated to the folder in which you wish to read and write files, select "More" and then "Set As Working Directory".

Plots: Plots The Plots screen displays all of your plots. There are buttons for opening the plot in a separate window and exporting the plot as a pdf or jpeg (though you can also do this with code using the pdf() or jpeg() functions.)

Packages: Displays a list of all the R packages that have been installed on your hard disc and indicates whether or not they are currently loaded. The word "packages" refers to the information that is displayed. Packages that have been installed but have

not yet been loaded are left unchecked, while those that are loaded during the current session are indicated with a checkmark.

Help -- Menu for getting assistance using R functions. You have the option of typing the name of a function into the search field, or you may look for a function in the code that matches the name.

What Sets the Environment Apart from the Rest of the List?

- Everything in a setting has a designated name.
- The environment has a parent environment.
- Environments adhere to the semantics of the reference.

Create a New Environment

Using the `new.env()` method in R programming allows for the creation of a new environment. In addition, you may access the variables by using the `$` symbol or the `[[]]` operator. Yet, each variable is kept track of in its own distinct memory address. There are four unique environments, which are denoted by the functions `globalenv()`, `baseenv()`, `emptyenv()` and `environment ()`.

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

Parameters:

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

- **Data Types**

While conducting programming in any programming language, you need to utilise numerous variables to store various information. Memory areas that are set aside specifically for the purpose of storing values are what are known as variables. When you create a variable, you are effectively reserving some space in the memory of the computer.

You could find it useful to store information using a variety of data types, such as character, wide character, integer, floating point, double floating point, Boolean and so on. Memory is allotted and managed by the operating system according to the data type of a variable. The operating system also determines what data may be saved in the memory that has been reserved.

R, in contrast to other programming languages such as C and Java, does not need its variables to be declared as belonging to a particular data type. R-Objects are used to assign values to variables and the data type of the R-object is then taken and used as the variable's data type. R objects come in a wide variety of flavours. The ones that are most often utilised are:

- Vectors
- Lists
- Matrices
- Arrays
- Factors
- Data Frames

There are six different data types of these atomic vectors, which are also referred

Notes

to as the six different classes of vectors. The vector object is the simplest of these objects. The atomic vectors serve as the foundation upon which the other R-Objects are constructed.

Data Type	Example	Verify
Logical	TRUE, FALSE	Live Demo v <- TRUE print(class(v)) it produces the following result – [1] "logical"
Numeric	12.3, 5, 999	Live Demo v <- 23.5 print(class(v)) it produces the following result – [1] "numeric"
Integer	2L, 34L, 0L	Live Demo v <- 2L print(class(v)) it produces the following result – [1] "integer"
Complex	3 + 2i	Live Demo v <- 2+5i print(class(v)) it produces the following result – [1] "complex"
Character	'a', "good", "TRUE", '23.4'	Live Demo v <- "TRUE" print(class(v)) it produces the following result – [1] "character"
Raw	"Hello" is stored as 48 65 6c 6c 6f	Live Demo v <- charToRaw("Hello") print(class(v)) it produces the following result – [1] "raw"

In R programming, the most fundamental data types are R-objects known as vectors. These vectors may carry items of a variety of different classes, as was seen before. Please take note that the six categories of classes are not the only possible types of classes in R. For instance, we might make use of a number of atomic vectors in order to construct an array whose class would be array.

Vectors

Use the `c()` function, which means to combine the components into a vector, when you want to construct a vector that consists of more than one element. This is the case when you want to generate a vector.

Notes

```
# Create a vector.
apple <- c('red','green',"yellow")
print(apple)

# Get the class of the vector.
print(class(apple))
```

The following is the outcome that we get when we run the code in the previous sentence:

```
[1] "red"  "green" "yellow"
[1] "character"
```

Lists

A list is an R-object which can contain many different types of elements inside it like vectors, functions and even another list inside it.

```
# Create a list.
list1 <- list(c(2,5,3),21.3,sin)

# Print the list.
print(list1)
```

When we execute the above code, it produces the following result –

```
[[1]]
[1] 2 5 3

[[2]]
[1] 21.3

[[3]]
function (x) .Primitive("sin")
```

Matrices

A matrix is a two-dimensional rectangular data collection. It is possible to generate it by providing a vector value as an input to the matrix function.

```
# Create a matrix.
M = matrix( c('a','a','b','c','b','a'), nrow = 2, ncol = 3, byrow = TRUE)
print(M)
```

The following is the outcome that we get when we run the code in the previous sentence:

```
[,1] [,2] [,3]
[1,] "a"  "a"  "b"
[2,] "c"  "b"  "a"
```

Notes

Arrays

Arrays can have any number of dimensions, in contrast to matrices, which are only allowed to have two dimensions. The dim property, which is sent into the array method, is responsible for generating the necessary number of dimensions. In the following illustration, we will build an array with two components, each of which will be a 3x3 matrix.

```
# Create an array.
a <- array(c('green','yellow'),dim = c(3,3,2))
print(a)
```

The following is the outcome that we get when we run the code in the previous sentence:

```
, , 1

[1,] [2] [3]
[1,] "green" "yellow" "green"
[2,] "yellow" "green" "yellow"
[3,] "green" "yellow" "green"

, , 2

[1,] [2] [3]
[1,] "yellow" "green" "yellow"
[2,] "green" "yellow" "green"
[3,] "yellow" "green" "yellow"
```

Factors

The R-objects that are produced by utilising a vector are referred to as factors. It stores the vector in addition to the labels that correspond to the individual values of the elements in the vector. The labels are always character, regardless of whether the input vector contains numeric values, character values, Boolean values, or anything else. They are helpful in constructing models based on statistics.

The factor() function is responsible for the creation of factors. The count of levels may be obtained with the nlevels function.

```
# Create a vector.
apple_colours <- c('green','green','yellow','red','red','red','green')

# Create a factor object.
factor_apple <- factor(apple_colours)

# Print the factor.
print(factor_apple)
print(nlevels(factor_apple))
```

The following is the outcome that we get when we run the code in the previous sentence:

```
[1] green green yellow red red red green
Levels: green red yellow
[1] 3
```

Notes

Data Frames

Data frames are tabular data objects. As contrast to a matrix, a data frame allows multiple types of data to be included in each column. The first column may include numeric information, the second column may contain character information and the third column may contain logical information. It is a list consisting of vectors that are all the same length.

The **data.frame()** method is what's responsible for creating data frames.

```
# Create the data frame.
BMI <- data.frame(
  gender = c("Male", "Male", "Female"),
  height = c(152, 171.5, 165),
  weight = c(81,93, 78),
  Age = c(42,38,26)
)
print(BMI)
```

The following is the outcome that we get when we run the code in the previous sentence:

	gender	height	weight	Age
1	Male	152.0	81	42
2	Male	171.5	93	38
3	Female	165.0	78	26

- **Functions**

When you need to carry out a particular operation several times, functions might be quite helpful. A function is able to create output by executing acceptable R commands that are included within the function itself. A function is able to receive parameters as input. When constructing a function in the R Programming Language, the function name and the file in which the function is being created do not need to be the same. Moreover, you can have one or more function definitions in a single R file.

Types of Function in R Language

- **Built-in Functions:** The built-in functions of R include `sq()`, `mean()` and `max()`; users may directly call these functions within the application.
- **User-defined Functions:** The R programming language gives us the ability to create our own functions.

Function Definition

Using the keyword `function` will result in the creation of a R function. The following is an example of the fundamental syntax of a R function definition:-

Notes

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

Function Components

The following are the many components that make up a function:

- **Function Name** — The actual name of the function is contained inside this field. It is kept in the R environment as an object with this name and this storage location.

```
Function_name <- function(arguments){
  function_body
  return (return)
}
```

Where `function_name` is the name of the function,
 arguments are the input arguments needed by the function,
`function_body` is the body of the function,
`return` is the return value of the function.

- **Arguments** — A placeholder is referred to as an argument. When you call a function, you give a value to the argument that the function receives. It is possible for a function to have no parameters at all since arguments are not required. Additional parameters can have default values.

```
function_name(arguments)
```

A function must always be called with the appropriate amount of parameters, as this is the default behaviour. This means that if your function requires two parameters, you must provide in exactly two arguments when you call the function; you cannot pass in more or less arguments.

Example

This function expects 2 arguments, and gets 2 arguments:

```
my_function <- function(fname, lname) {
  paste(fname, lname)
}
my_function("Peter", "Griffin")
```

- **Function Body** - The “function body” section of a function is where you’ll find a collection of statements that outline the purpose of the function.

The syntax for getting the body of a function using the `body()` function:

```
body(fun = sys.function(sys.parent()))
```

The syntax for setting the body of a function:

```
body(fun, envir = environment(fun)) <- value
```

Parameters

Notes

The body() function takes the following parameter values:

fun: This represents the function object.

envir: This represents the environment in which the function should be defined.

value: This represents the value to make up the value of the function.

- Return Value - A function's return value is the final expression in the function body that is evaluated. It is referred to as the "return value."

To let a function return a result, use the return() function:

Example

```
my_function <- function(x) {
  return (5 * x)
}

print(my_function(3))
print(my_function(5))
print(my_function(9))
```

The output of the code above will be:

```
[1] 15
[1] 25
[1] 45
```

R comes with a plethora of pre-defined functions that may be easily invoked within a programme without having to first define them. We also can build and utilise what are known as user defined functions for our own purposes.

Built-in Function

A few instances of built-in functions are the seq(), mean(), max(), sum(x) and paste(...) commands, amongst others. They receive direct calls from programmes that have been built by users.

```
# Create a sequence of numbers from 32 to 44.
print(seq(32,44))

# Find mean of numbers from 25 to 82.
print(mean(25:82))

# Find sum of numbers from 41 to 68.
print(sum(41:68))
```

The following is the outcome that we get when we run the code in the previous sentence:

```
[1] 32 33 34 35 36 37 38 39 40 41 42 43 44
[1] 53.5
[1] 1526
```

Notes

User-defined Function

In R, it is possible to build user-defined functions. They are tailored to exactly what the user requires and once they have been made, they may be utilised in the same way as the built-in functions. An illustration of how a function may be developed and put into use can be found below.

```
new.function <- function(a) {
  for(i in 1:a) {
    b <- i^2
    print(b)
  }
}
```

- **Loops in R (for, while, repeat)**

To repeat the execution of a section of code more than once, the R programming language requires a control structure. Programming principles that are considered to be among the most fundamental and reliable include loops. A control statement known as a loop enables several iterations of a single statement or a series of statements at the same time. Iterating or cycling through a process is what the term “looping” refers to.

Within its framework, a loop poses a question to be answered. If the response to that question demands a certain action, then that action will be carried out. The identical question is posed in several iterations until some additional action is made. An instance of the question being posed within the framework of the loop is referred to as a “iteration” of the loop. The control statement and the loop body are the two parts that make up the entirety of a loop. The control statement is what decides which statements get run based on the condition and the loop body is where the collection of statements that are going to get executed is stored.

A simple loop is all that is required for a programmer to be able to utilise a programme to achieve the desired effect of repeatedly executing the same lines of code.

In R programming, there are three different forms of loops:

- For Loop
- While Loop
- Repeat Loop

For Loop in R

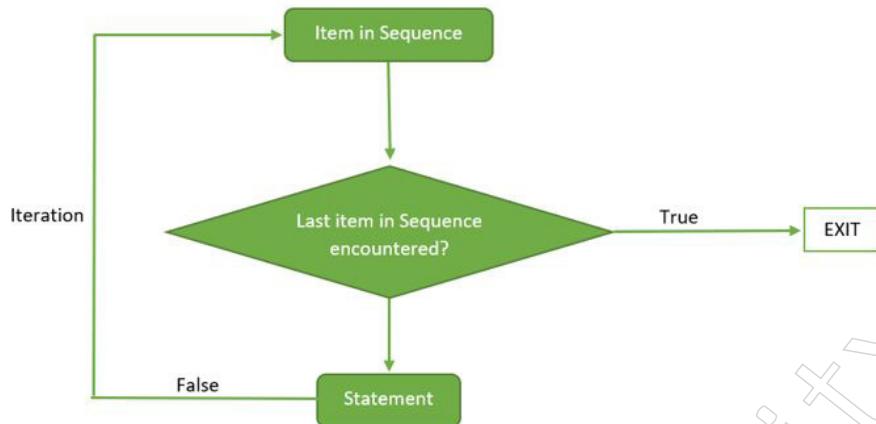
It is a sort of control statement that makes it simple to build a loop that must execute a set of statements or a set of statements many times. This type of loop must run statements or sets of statements. It is usual practise to use the for loop to iterate through the elements in a series. It is an entry-controlled loop, which means that, within this loop, the test condition is evaluated first and only then is the body of the loop run. If the test condition is found to be false, the loop body will not be executed.

R – For loop Syntax:

```
for (value in sequence)
{
```

```
statement
}
```

For Loop Flow Diagram:



The programmes listed below illustrate the use of the for loop in R programming.

Example: R programme that displays the numbers 1 through 5 using a for loop.

```
# R program to demonstrate the use of for loop

# using for loop

for (val in 1: 5)
{
  # statement
  print(val)
}
```

Output:

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
```

Here, the for loop iterates over a sequence containing the integers 1 through 5. Each item in the sequence is displayed during each iteration.

While Loop in R

It is a form of control statement that will repeatedly execute a statement or set of statements until the specified condition becomes false. It is also an entry-controlled loop in which the test condition is evaluated before the loop body is executed; if the test condition is false, the loop body is not executed.

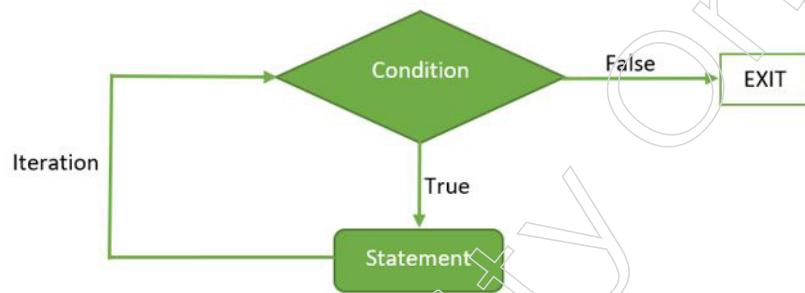
R – While loop Syntax:

Notes

Notes

```
while ( condition )
{
    statement
}
```

While loop Flow Diagram:



The programmes listed below illustrate the while loop in R programming.

Example: R code to display the numerals 1 through 5 using a while loop.

```
# R program to demonstrate the use of while loop

val = 1

# using while loop
while (val <= 5)
{
    # statements
    print(val)
    val = val + 1
}
```

Output:

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
```

The initial value of the variable is set to 1. In each iteration of the while loop, the condition is evaluated and the value of val is displayed; val is then incremented until it reaches 5 and the condition becomes false, at which point the loop exits or ends.

Repeat Loop in R

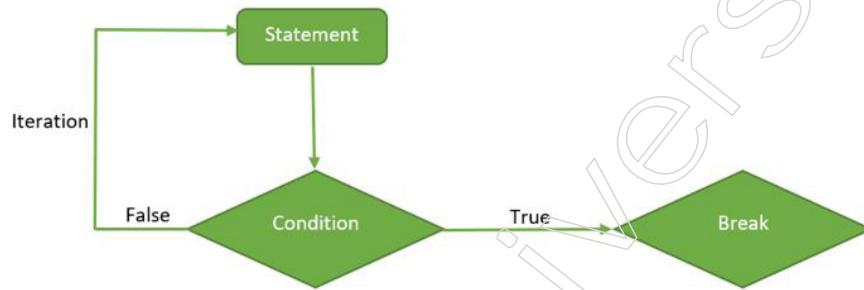
It is a simple loop that repeatedly executes the same statement or group of statements until the stop condition is met. Repeat loop lacks a condition to terminate the loop; a programmer must place a condition within the loop's body and declare a

break statement to terminate this loop. If there is no condition in the body of the repeat loop, then it will repeat indefinitely.

R – Repeat loop Syntax:

```
repeat
{
  statement
  if( condition )
  {
    break
  }
}
```

Repeat loop Flow Diagram:



The break keyword is used as a leap statement to terminate the repeat cycle. The programmes listed below demonstrate the use of repeat loops in R programming.

Example: R programme to display the numerals 1 through 5 using a repeat cycle.

```
# R program to demonstrate the use of repeat loop
val = 1
# using repeat loop
repeat
{
  # statements
  print(val)
  val = val + 1

  # checking stop condition
  if(val > 5)
  {
```

```
# using break statement
  # to terminate the loop
  break
}
```

Notes

Notes

Output:

```
[1] 1  
[1] 2  
[1] 3  
[1] 4  
[1] 5
```

In the preceding programme, the variable val is initialised to 1, and its value is displayed in each iteration of the repeat loop before being incremented until it becomes greater than 5. If the value of val exceeds 5, the break statement is executed to terminate the loop.

- **Data Structures**

A data structure is a specific method of arranging data in a computer so that it may be utilised in the most efficient manner possible. The goal is to simplify a variety of processes by lessening the demands placed on both space and time. In the programming language R, data structures are tools that may contain several values at once. R's fundamental data structures are often classified according to their dimensions (one-dimensional, two-dimensional, or more) as well as whether or not they are homogeneous (meaning that every element must be of the same type) or heterogeneous (the elements are often of various types). As a result of this, there are six different categories of data that are employed in data analysis the majority of the time.

The following are some of the most important data structures that are utilised in R:

- Vectors
- Lists
- Dataframes
- Matrices
- Arrays
- Factors

1.4.2 R -Packages, Dataset Reading, Programming, Statistical Introduction

- **R Packages**

The package is an effective method for arranging the work in an orderly fashion and communicating its contents to others. In most cases, a package will consist of the following components: code (and not only R code, either!), documentation for the package and the functions included therein some tests to ensure that everything operates as it should and data sets.

Packages in R

In the R programming language, a package is a collection of R functions, code that has been produced and example data. Within the R environment, these are all kept in

a directory that is referred to as the “library.” During the installation process, R will, by default, install a collection of packages. Once we have started the R console, the only packages that will be available will be the default ones. In order for other packages that have previously been installed to be utilised by the R application that’s getting to use them, those other packages need to be loaded specifically.



Tidyr	The term tidy is derived from tidy, which means clear. Thus, the tidy utility is used to organise the data. This package complements dplyr well. This programme represents a development of the reshape2 utility.
ggplot2	R enables declarative graphic creation. For this, R provides the ggplot package. This package's refined and high-quality graphs distinguish it from other visualisation packages.
Ggraph	ggraph is an extension of ggplot provided by R. ggraph eliminates the limitation of ggplot dependence on tabular data.
dplyr	R enables us to conduct data manipulation and analysis. The dplyr library is provided by R for this purpose. This library provides multiple functions for the R data frame.
tidyquant	The tidyquant is a financial software programme used for quantitative financial analysis. This package introduces a financial package to the tidyverse universe that is used for integrating, analysing, and visualising data.
dygraphs	The dygraphs package offers a charting interface to the primary JavaScript library. This R package is primarily used to illustrate time-series data.
leaflet	R provides the leaflet utility for interactively creating visualisations. This package is an open-source for JavaScript library. Popular websites such as the New York Times, Github, and Flicker, among others, use leaflet. The leaflet package makes it easier to interact with these sites.
ggmap	For delineating spatial visualization, the ggmap package is used. It is a mapping application containing numerous geolocating and routing utilities.

Notes

glue	R provides the glue package required to perform data wrangling operations. This package is utilised to evaluate R expressions contained within a string.
shiny	R enables the creation of interactive and aesthetically appealing web applications by delivering a shiny package. This package contains a variety of elements for HTML, CSS, and JavaScript.
plotly	The plotly programme offers interactive and high-quality graphs online. This module extends the -plotly.js JavaScript library.
tidytext	The tidytext package provides various text mining functions for word processing and conducting analysis with ggplot, dplyr, and other tools.
stringr	The stringr package provides containers for the 'stringi' package that are simple and consistent to use. The stringi programme simplifies standard string operations.
reshape2	Using the melt () and decast () functions, this package enables flexible data reorganisation and aggregation.
dichromat	The R dichromat programme is utilised to eliminate Red-Green or Blue-Green colour contrasts.
digest	The digest package is used for the creation of cryptographic hash objects of R functions.
MASS	The MASS utility contains numerous statistical functions. It contains datasets that correspond to the text "Modern Applied Statistics with S."
caret	The caret package in R allows us to execute classification and regression tasks. CaretEnsemble is a caret feature that enables the combination of multiple models.
e1071	The e1071 library provides essential functions for data analysis, such as Naive Bayes, Fourier Transforms, SVMs, and Clustering, among other functions.
sentimentr	The sentiment package contains functions for sentiment analysis. It is utilised for calculating the polarity of text at the sentence level and for aggregating rows or grouping variables.

Install an R-Packages

There are numerous methods to deploy R Package, including the following:

- **Installing Packages From CRAN:** Installing a package from CRAN requires the package's name and the following command:

```
install.packages("package name")
```
- Installing a package from CRAN is the most common and straightforward method, as it requires only a single command. To install multiple packages simultaneously, it is sufficient to specify them as a character vector in the `install.packages()` function's first argument:

Example:

```
install.packages(c("vioplot", "MASS"))
```

- **Installing Bioconductor Packages:** In Bioconductor, the standard way to install a package is by first executing the following script:

```
source("https://bioconductor.org/biocLite.R")
```

- This will install essential functions for installing bioconductor packages, including the biocLite() function. To install the Bioconductor core packages, simply type it without any additional arguments:
- ```
biocLite()
```
- Type the package names explicitly as a character vector if we only want a few specific packages from this repository.

**Example:**

```
biocLite(c("GenomicFeatures", "AnnotationDbi"))
```

**Update, Remove and Check Installed Packages in R**

**To check what packages are installed on your computer, type this command:**

```
installed.packages()
```

**To update all the packages, type this command:**

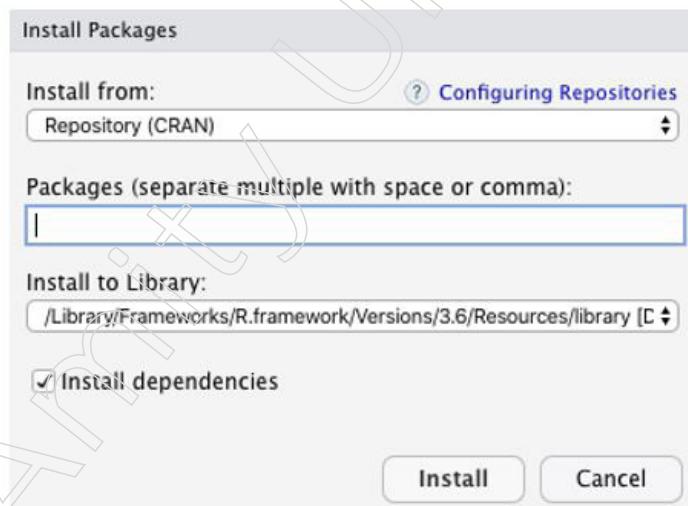
```
update.packages()
```

**To update a specific package, type this command:**

```
install.packages("PACKAGE NAME")
```

**Installing Packages Using RStudio UI**

In R Studio, navigate to Tools -> Install Package to bring up a pop-up window where you can enter the package you wish to install:



Under Packages, type, and search Package which we wish to install and then click on install button.

#### How to Load Packages in R Programming Language

When a package is installed, its features are immediately accessible. If only a few

## Notes

functions or data within a package are required on occasion, we can access them using the following notation.

```
packagename::functionname()
```

Example: Let's access the `births` function of the `babynames` module as an example. Then enter the following command:

```
babynames::births
```

### Output:

```
Console Terminal Jobs
~/
> babynames::births
A tibble: 109 x 2
 year births
 <int> <int>
1 1909 2718000
2 1910 2770000
3 1911 2809000
4 1912 2840000
5 1913 2869000
6 1914 2966000
7 1915 2965000
8 1916 2964000
9 1917 2944000
10 1918 2948000
... with 99 more rows
```

## What are Repositories?

You can retrieve and install packages from a repository since it is a central storage location for the packages themselves. Each organisation and developer normally maintains their own local repository, which is often hosted online and is available to users worldwide. The following list includes some of the most prominent R package repositories:

- **CRAN:** The Comprehensive R Archive Network (CRAN) is the official repository. CRAN is a network of web servers and ftp servers that are maintained by members of the R community from all around the world. It is coordinated by the R community and in order for a package to be published in CRAN, it must first pass a number of tests designed to guarantee that it complies with the policies established by CRAN.
- **Bioconductor:** It is a repository that is geared at open-source software that is used for bioinformatics. Bioconductor is a topic-specific repository. In a manner analogous to that of CRAN, it possesses its own submission and review procedures and its community is quite active, holding a number of conferences and meetings on an annual basis in order to ensure the highest possible level of quality.
- **Github:** It is currently the most widely used repository for open-source projects. The infinite storage space for open source, the integration with git, which is software for version control and the simplicity with which it is possible to share and interact with other people are all factors that contribute to its popularity.

## Install an R-Packages

There are many different approaches of installing R Package, some of them are as follows:

- **Installing Packages From CRAN:** We will require the package's name in order to install it from CRAN and then we will run the following command:

```
install.packages("package name")
```

Installing a package from CRAN is by far the most frequent and straightforward method, as it only requires the execution of a single command. To install more than one package simultaneously, all that is required of us is to enter the package names into the `install.packages()` function as a character vector in the first argument::

### Example:

```
install.packages(c("vioplot", "MASS"))
```

- **Installing Bioconductor Packages:** The following script must be run in order to install a package using the default method in Bioconductor, which is as follows:

```
source("https://bioconductor.org/biocLite.R")
```

This will install certain fundamental functions that are required in order to install bioconductor packages. One of these functions, known as `biocLite()`, is an example. Just typing it in without any other parameters will result in the installation of Bioconductor's core packages:

```
biocLite()
```

If there are only a few specific packages from this repository that we are interested in, we may enter their names directly as a character vector:

### Example:

```
biocLite(c("GenomicFeatures", "AnnotationDbi"))
```

## Update, Remove and Check Installed Packages in R

- Type this command into your computer's prompt to get a list of all the packages that have been installed on it:

```
installed.packages()
```

- To bring all the packages up to date, enter the following command:

```
update.packages()
```

- To bring an individual package up to date, run the following command:

```
install.packages ("PACKAGE NAME")
```

## Dataset Reading

A dataset in R is described as a central area within the package in RStudio where data from diverse sources are saved, maintained and made available for usage. Specifically, a dataset in RStudio is referred to as a dataset in R. In this day

## Notes

and age of big data, it has always been difficult to locate data that is not only clean and dependable, but also includes metadata that is simple and straightforward to understand. RStudio is an Integrated Development Environment (IDE) that gives programmers the ability to construct statistical models for use in graphics and statistical computing. The RStudio application, which provides the requisite usability for the specified use case, stores datasets in the R programming language in a format that is compatible with that application. One of the formats that may be purchased is called RStudio Desktop, while the other is called RStudio Server. Both are accessible on the market. The description of the dataset, on the other hand, does not make any assumptions about the file format and is therefore acceptable for any version.

### Using R-Studio

The following methods will be used to import data through R studio.

#### Steps:

- From the Environment tab, select Import Dataset from the menu
- Choose the file extension from the option
- In the third stage, input the filename or browse the desktop using the pop-up box.
- The selected file's dimensions will be displayed in a new window.
- Type the filename to view the output on the console.

#### Example:

```
display the dataset
Dataset
```

#### Output:

- The attach command is used to load the data onto the console for use.

#### Example:

```
To load the data for use
attach(dataset)
```

### How to Read DataSet into R?

There are two possible categories of the dataset and each form has its own particular approach to interpreting the dataset. The first type of dataset is one that has already been compiled and saved within an RStudio package, which the programmer is able to use directly. On the other hand, there is a second type of dataset that can be present in its raw form, which is denoted by the notation viz. excel, csv, database etc. In this section, we will investigate each of the distinct paths one at a time. We will look at a limited number of examples based on the dataset that is included in the RStudio package; however, we will not limit ourselves to the dataset itself as a topic for our discussion. In essence, we will investigate datasets that are geared specifically at the challenge of classification and regressions on their own.

### From the pre-defined dataset in the package:

The majority of the datasets may be accessed immediately by using the RStudio package that is located in the repository with the name “UCI Machine Learning.” The following qualities contribute to the widespread usage of these datasets, which in turn helps to explain their widespread popularity:

- The dataset is available for quick download.
- The datasets are quite tiny and as a result, they may be stored in memory.
- The datasets have been cleaned up for the most part; as a result, the step of cleaning the data may be skipped and one can proceed straight to the step of running the algorithms on the datasets.

Through the Comprehensive R Archive Network (CRAN) bridge, which enables third party libraries to download and keep the modules stored in the RStudio package, these packages are already in place, which enables developers to easily download and use them in their projects. This is made possible by the fact that these packages are already in place.

### Preliminary tasks

**Start RStudio as outlined here:** Launching RStudio and configuring the working directory

#### List of pre-loaded data

To view the pre-loaded data list, enter the **function data()**:

```
data()
```

The following is the output:

The screenshot shows the RStudio interface with the following components visible:

- Environment pane:** Shows "Global Environment" with "Environment is empty".
- Plots pane:** Displays a boxplot titled "Experiment" with four groups labeled A, B, D, and F. The y-axis ranges from 0 to 20. Group A has a median around 18, group B around 15, group D around 5, and group F around 18.
- Console pane:** Shows the R code being run:
 

```
> data()
>
> boxplot(count ~ spray, data = InsectSprays, col = "lightgray")
> # *add* notches (somewhat funny here):
> boxplot(count ~ spray, data = InsectSprays,
+ notch = TRUE, add = TRUE, col = "blue")
Warning message:
In bxp(list(stats = c(7, 11, 14, 18.5, 23, 7, 12, 16.5, 18, 2
1, :
 Quelques indentations ("notches") dépassent des jointures (
 "hinges") ('box') : utilisez peut-être notch=FALSE
> data()
```

## Notes

Loading a built-in R data

Load and print mtcars data as follow:

```
Loading
data(mtcars)

Print the first 6 rows

head(mtcars, 6)
```

|                   | mpg  | cyl | disp | hp  | drat | wt    | qsec  | vs | am | gear | carb |
|-------------------|------|-----|------|-----|------|-------|-------|----|----|------|------|
| Mazda RX4         | 21.0 | 6   | 160  | 110 | 3.90 | 2.620 | 16.46 | 0  | 1  | 4    | 4    |
| Mazda RX4 Wag     | 21.0 | 6   | 160  | 110 | 3.90 | 2.875 | 17.02 | 0  | 1  | 4    | 4    |
| Datsun 710        | 22.8 | 4   | 108  | 93  | 3.85 | 2.320 | 18.61 | 1  | 1  | 4    | 1    |
| Hornet 4 Drive    | 21.4 | 6   | 258  | 110 | 3.08 | 3.215 | 19.44 | 1  | 0  | 3    | 1    |
| Hornet Sportabout | 18.7 | 8   | 360  | 175 | 3.15 | 3.440 | 17.02 | 0  | 0  | 3    | 2    |
| Valiant           | 18.1 | 6   | 225  | 105 | 2.76 | 3.460 | 20.22 | 1  | 0  | 3    | 1    |

### Most used R built-in data sets

#### mtcars: Motor Trend Car Road Tests

The data was extracted from the US journal Motor Trend in 1974 and includes fuel utilisation and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

- View the content of mtcars data set:

```
1. Loading
data("mtcars")

2. Print
head(mtcars)
```

- It contains 32 observations and 11 variables:

```
Number of rows (observations)
nrow(mtcars)
[1] 32

Number of columns (variables)
ncol(mtcars)
[1] 11
```

- Description of variables:
  1. mpg: Miles/(US) gallon
  2. cyl: Number of cylinders
  3. disp: Displacement (cu.in.)
  4. hp: Gross horsepower
  5. drat: Rear axle ratio
  6. wt: Weight (1000 lbs)
  7. qsec: 1/4 mile time

8. vs: V/S
9. am: Transmission (0 = automatic, 1 = manual)
10. gear: Number of forward gears
11. carb: Number of carburetors

**Notes**

## Iris

The iris data set provides the sepal length, sepal width, petal length, and petal width, in centimetres, for 50 flowers from each of the three species of iris. Iris setosa, versicolor, and virginica are the species.

```
data("iris")
head(iris)

Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1 5.1 3.5 1.4 0.2 setosa
2 4.9 3.0 1.4 0.2 setosa
3 4.7 3.2 1.3 0.2 setosa
4 4.6 3.1 1.5 0.2 setosa
5 5.0 3.6 1.4 0.2 setosa
6 5.4 3.9 1.7 0.4 setosa
```

Let us take a look at some of the datasets that are most well-known in the field of data science practitioners.

### 1. Dataset Library

There is no need to load the library since the components that it contains are already included in the default installation of RStudio; hence, loading the library is not required. This package comes with a variety of libraries already installed on your computer. Executing the following command is one of the ways you are able to look at the many datasets that are accessible in this library.

#### Code:

```
library(help = "datasets")
```

### 2. Iris Dataset

This dataset includes the numerous varieties of Iris flowers that may be determined based on the various feature sets and measurements of the blooms. There are three different categories of variations and each one is defined by a set of four characteristics: the length of the sepal, the breadth of the sepal, the length of the petal and the width of the petal. Using the following command will load the dataset into memory for you to work with.

#### Code:

```
data(iris)
```

This data is utilised extensively in the testing of algorithms that are geared towards the category of problems known as multi-class classification issues.

## Notes

### 3. Longley's Economic Dataset

On the basis of the many different economic indicators, this dataset includes the percentage of the population that was gainfully employed for a specific year. There are six distinct characteristics that explain the percentage of individuals who are employed, which is displayed in the column titled "Employed." In the future, one may forecast the percentage of people who might be employed based on the economic indicators in a certain year. Using the following command will load the dataset into memory for you to work with.

#### Code:

```
data(longley)
```

These data are utilised extensively in the process of testing algorithms that are specific to the category of regression problems.

### 4. mlbench Library

This collection contains data pertaining to a variety of real-world benchmark challenges from around the globe. Executing the command will result in the installation of the library.

#### Code:

```
install.packages("mlbench")
```

Executing the command will result in the library being loaded into memory.

#### Code:

```
library(mlbench)
```

Executing the following code will, in a manner analogous to that of the datasets library, return a list of all the datasets contained inside the mlbench library.

#### Code:

```
library(help = "mlbench")
```

### Functions for Reading Data into R:

There are a few very useful functions for reading data into R.

#### 1.

| Functions                            | Uses                                                                                     |
|--------------------------------------|------------------------------------------------------------------------------------------|
| read.table() and read.csv() function | These are two of the most common ones used to read tabular data into the programme.      |
| readLines()                          | It is the function that is called when lines need to be read from a text file.           |
| source() function                    | For reading code files from one R programme into another R programme, it is very useful. |
| Dget() function                      | It is utilised in the reading in of R code files.                                        |

|                        |                                                                                   |
|------------------------|-----------------------------------------------------------------------------------|
| load() function        | It is what's responsible for reading in previously stored workspaces.             |
| unserialize() function | It is utilised for reading individual R objects that are stored in binary format. |

**Notes**

- **Programming**

R is a programming language that also serves as an environment for analysing data and doing statistical computations. R is a programming language that was developed at the University of Auckland in New Zealand by Ross Ihaka and Robert Gentleman. R is a freely distributable programming language that is available for usage on a variety of systems, including but not limited to Windows, Linux and Mac. It often includes a command-line interface and gives a comprehensive list of programmes that may be used to carry out various operations. R is a procedural and object-oriented programming language that may be interpreted. It supports both types of programming. With the backing of over 10,000 and counting free packages in the CRAN library, R has quickly become the most popular language used for statistical computing and data analysis.

### Syntax of R program

Variables, Comments and Keywords are the three components that make up a R programme. The data may be stored in variables, comments can improve the readability of the code and keywords are reserved words that have a special meaning to the compiler.

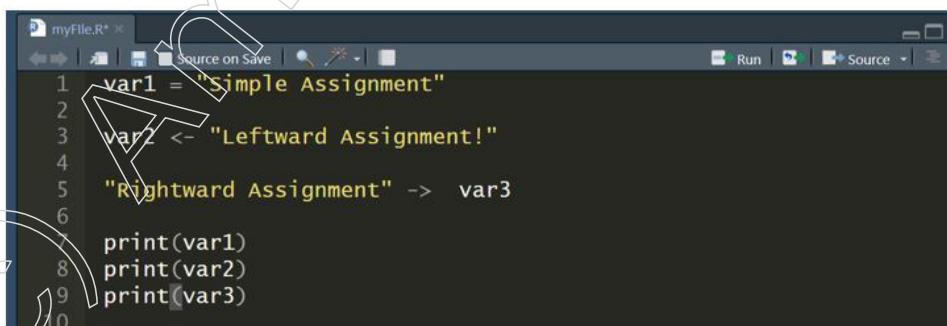
- **Variables in R**

In the past, we wrote all of our code inside of a print() function, but we do not currently have a method to address them in order to carry out further activities. This issue may be resolved by making use of variables, which, just like in any other programming language, are the names given to memory places that are designated specifically for the purpose of storing data of any kind.

There are three ways in which the assignment can be denoted in R:

1. = (Simple Assignment)
2. <- (Leftward Assignment)
3. -> (Rightward Assignment)

### Example:



```

myFile.R*
1 var1 = "Simple Assignment"
2
3 var2 <- "Leftward Assignment!"
4
5 "Rightward Assignment" -> var3
6
7 print(var1)
8 print(var2)
9 print(var3)
0

```

## Notes

### Output:

“Simple Assignment”

“Leftward Assignment!”

“Rightward Assignment”

- **Comments in R**

Your code's readability can be improved by the inclusion of comments, which are exclusively intended for the user and are thus ignored by the interpreter. Although R only supports single-line comments, it is possible to utilise multiline comments by employing a straightforward workaround, which will be explained in more detail below. Comments on a single line can be written by inserting a hash symbol (#) at the beginning of the sentence.

### Example:

```

myFile.R* <-->
1 # This is a single line comment
2 print("This is fun!")
3
4 if(FALSE)
5 {
6 "This is multi-line comment which should be put inside either a
7 single or a double quote"
8 }
9

```

### Output:

[1] “This is fun!”

From the above output, we can see that both comments were ignored by the interpreter.

- **Keywords in R**

Due to the unique significance attached to each word, a computer programme will not allow a keyword to be used anywhere else in the code, including as the name of a variable, a function, or anything else.

We can view these keywords by using either `help(reserved)` or `?reserved`.

#### Reserved words in R

|             |          |             |               |      |
|-------------|----------|-------------|---------------|------|
| if          | else     | while       | repeat        | for  |
| function    | in       | next        | break         | TRUE |
| FALSE       | NULL     | Inf         | NaN           | NA   |
| NA_integer_ | NA_real_ | NA_complex_ | NA_character_ | ...  |

- Control-flow statements and user-defined functions can be declared using the keywords if, else, repeat, while, function, for, in, next and break. Other control-flow statements include repeat, while and function.
- The ones that are still around are the ones that are utilised as constants, such as how TRUE and FALSE are used as boolean constants.

- The value Not a Number is specified by the NaN notation and the NULL notation is used to describe an undefined value.
- The value "inf" denotes an infinite amount.

**Note:** Take note that R is a language that pays attention to case, thus "TRUE" and "True" are not the same thing.

#### • Statistical Introduction

The collecting of data, its organisation, analysis, interpretation and presentation are the primary focuses of the statistical method, which is a subfield of mathematical analysis. The statistical analysis enables a better utilisation of the large amounts of data that are accessible and increases the overall efficacy of the solutions.

### R – Statistics

R is a computer language that is utilised for statistical computation and visuals in the field of the environment. The following is an introduction to several fundamental concepts in statistics, such as the normal distribution (also known as a bell curve), central tendency (the mean, median and mode), variability (25%, 50% and 75% quartiles), variance, standard deviation, modality and skewness.

### Data Concepts

Before we can begin to understand the ideas of statistics, we need to be familiar with the many formats used for storing data. Data may be organised and presented in a variety of ways.

#### These are some formats:

- Vector
- Dataframe
- Variable
- Continuous Data
- Discrete Data
- Normal Data
- Categorical Data
- Normal Distribution
- Skewed Distribution

### Statistics in R

#### • Average, Variance and Standard Deviation in R

The R programming language is a type of open-source programming language that is frequently employed in the statistical software and data analysis tools industries. The Command-line interface is often included with R installations. R is a programming language that can run on a variety of operating systems, including Windows, Linux and macOS. Calculating the average, variance and standard deviation are all made very simple by the programming language R.

## Notes

### Average in R Programming

Average a number expressing the central or typical value in a set of data, in particular the mode, median, or (most commonly) the mean, which is calculated by dividing the sum of the values in the set by their number. The basic formula for the average of n numbers  $x_1, x_2, \dots, x_n$  is

$$A = (x_1 + x_2 + \dots + x_n) / n$$

### Variance in R Programming Language

The total of the squares of the discrepancies between all of the numbers and the means is the variance. The following is the mathematical formula for calculating variance:

$$\text{Formula : } \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

### Standard Deviation in R Programming Language

Standard Deviation is the square root of variance. It is a measure of the extent to which data varies from the mean. The mathematical formula for calculating standard deviation is as follows,

$$\text{Standard Deviation} = \sqrt{\text{variance}}$$

### Average in R Programming

Average a number expressing the central or typical value in a set of data, such as the mode, median, or (most frequently) the mean, which is calculated by dividing the sum of the values by the number of values. The basic formula for the average of n numbers  $x_1, x_2, \dots, x_n$  is

#### Example:

Suppose we have are 8 data points,

2, 4, 4, 4, 5, 5, 7, 9

The average of these 8 data points is,

$$A = \frac{2 + 4 + 4 + 4 + 5 + 5 + 7 + 9}{8} = 5$$

### Computing Average in R Programming

To compute the average of a set of values, R provides the `mean()` function. This function accepts a Numerical Vector as an argument and results in the average/mean of that Vector.

**Syntax:** `mean(x, na.rm)`

#### Parameters:

- `x`: Numeric Vector
- `na.rm`: Boolean value to ignore NA value

**Example 1:**

```
R program to get average of a list

Taking a list of elements
list = c(2, 4, 4, 4, 5, 5, 7, 9)

Calculating average using mean()
print(mean(list))
```

**Output:**

[1] 5

**Example 2:**

```
R program to get average of a list

Taking a list of elements
list = c(2, 40, 2, 502, 177, 7, 9)

Calculating average using mean()
print(mean(list))
```

**Output:**

[1] 105.5714

**Variance in R Programming Language**

The variance is the sum of the squares of the differences between all integers and their respective means. The formula for variance in mathematics is as follows:

$$\text{Formula : } \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

*where,*

$\mu$  is mean,

$N$  is the total number of elements or frequency of distribution.

**Example:**

Let us take at the same dataset that we have taken in average. Calculate first the deviations of each data point from the mean, then square each result.

$$\begin{aligned}
 (2 - 5)^2 &= (-3)^2 = 9 & (5 - 5)^2 &= 0^2 = 0 & [\text{Tex}] \text{variance} = \\
 (4 - 5)^2 &= (-1)^2 = 1 & (5 - 5)^2 &= 0^2 = 0 \\
 (4 - 5)^2 &= (-1)^2 = 1 & (7 - 5)^2 &= 2^2 = 4 \\
 (4 - 5)^2 &= (-1)^2 = 1 & (9 - 5)^2 &= 4^2 = 16. \\
 \frac{9 + 1 + 1 + 1 + 0 + 0 + 4 + 16}{8} &= 4 & & & [/tex]
 \end{aligned}$$

## Notes

### Computing Variance in R Programming

One can calculate the variance by using var() function in R.

**Syntax:** var(x)

**Parameters:**

*x: numeric vector*

#### Example 1:

```
R program to get variance of a list

Taking a list of elements
list = c(2, 4, 4, 4, 5, 5, 7, 9)

Calculating variance using var()
print(var(list))
```

#### Output:

[1] 4.571429

#### Example 2:

```
R program to get variance of a list

Taking a list of elements
list = c(212, 231, 234, 564, 235)

Calculating variance using var()
print(var(list))
```

#### Output:

[1] 22666.7

### Standard Deviation in R Programming Language

Standard Deviation is the square root of variance. It is a measurement of the deviation of data from the norm. Following is the mathematical formula for calculating standard deviation:

$$\text{Standard Deviation} = \sqrt{\text{variance}}$$

#### Example:

Standard Deviation for the above data,

$$\text{Standard Deviation} = \sqrt{4} = 2$$

### Computing Standard Deviation in R

Using the sd() function of R, the standard deviation can be calculated.

**Syntax:** `sd(x)`

**Parameters:**

**x:** numeric vector

#### Example 1:

```
R program to get
standard deviation of a list

Taking a list of elements
list = c(2, 4, 4, 4, 5, 5, 7, 9)

Calculating standard
deviation using sd()
print(sd(list))
```

#### Output:

```
[1] 2.13809
```

#### Example 2:

```
R program to get
standard deviation of a list

Taking a list of elements
list = c(290, 124, 127, 899)

Calculat
ing standard
deviation using sd()
print(sd(list))
```

#### Output:

```
[1] 367.6076
```

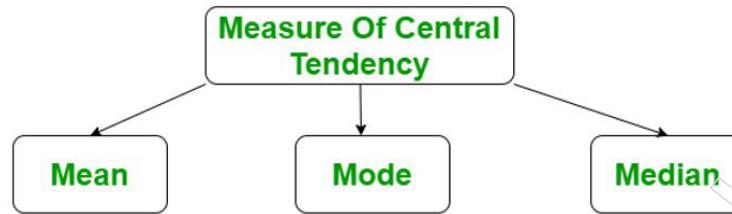
#### • Mean, Median and Mode in R Programming

The R programming language's implementation of the standard deviation

The square root of variance is the formula for calculating standard deviation. It is a measurement of how significantly the data deviates from the mean value. The following is a mathematical formula that may be used to calculate the standard deviation:

- Mean
- Median
- Mode

## Notes



### Prerequisite:

Before we can perform any kind of computation, we need to first of all prepare our data. We should save our data in separate files ending in.txt or.csv and it is recommended that you save the file in the directory that you are now working in. Following that import, your data will be in R formatted as follows:

### Mean in R Programming Language

It is calculated by taking the total number of observations and dividing it by the sum of all the observations. It is sometimes referred to as the average, which is calculated by dividing the total by the total count.

$$\text{Mean}(\bar{x}) = \frac{\sum x}{n}$$

Where, n = number of terms

### Median in R Programming Language

It is the value that is in the middle of the entire data set. The data is divided in half as a result of this operation. If the number of items in the data set is odd, then the element in the centre is considered to be the median; if the number of elements is even, then the median is determined by taking the average of the two elements in the centre.

|                   |                                |
|-------------------|--------------------------------|
| <u><b>Odd</b></u> | <u><b>Even</b></u>             |
| $\frac{n+1}{2}$   | $\frac{n}{2}, \frac{n}{2} + 1$ |

Where n = number of terms

**Syntax:** `median(x, na.rm = False)`

**Where,** X is a vector and na.rm is used to remove missing value

### Mode in R Programming Language

It is the value that appears in the data set the most frequently at this point in time. If the frequency of each data point is the same, then the data collection could not include a mode at all. In addition, there is the possibility that we might have more than one mode if we come across two or more data points that have the same frequency. As R does not have a mode-finding function as part of the standard distribution, we will need to either write our own mode-finding function or make use of a package called modeest.

### 1.4.3 Importance of Data Visualisation, Visualisation Aesthetics

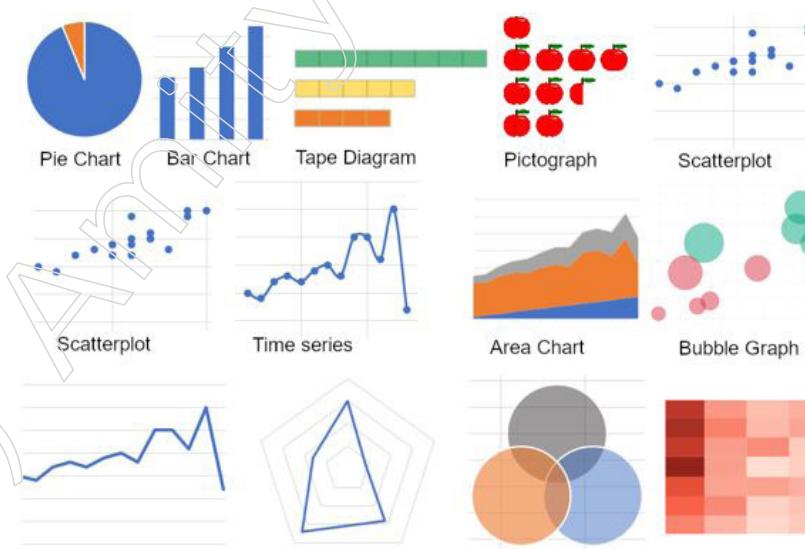
The practise of portraying data in a visual and understandable fashion with the

purpose of enhancing a user's ability to comprehend that data is known as information visualisation. Examples of information visualisation that are used often include dashboards and scatter plots. Users are able to get actionable insights from abstract data in a way that is both efficient and effective thanks to information visualisation, which depicts an overview and demonstrates key relationships.

The process of making data more easily edible and converting raw data into insights that can be acted upon is significantly aided by the practise of information visualisation. It draws inspiration from a variety of disciplines, including but not limited to human-computer interaction, graphic design, computer science and cognitive science. Representations in the manner of a globe map, line graphs and three-dimensional virtual building or town plan designs are a few examples.

Understanding the information requirements of the audience for whom the visualisation is intended is often the first step in the process of developing the visualisation. Finding out how, when and where the visualisation will be utilised may be uncovered using qualitative research methods like as user interviews. A designer can establish which kind of data organisation is required for the users' goals to be achieved by applying these insights to the design process. When the information has been structured in a manner that makes it easier for people to grasp it and makes it easier for them to apply it so that they may achieve their goals, the next tools that a designer will put out to use are visualisation approaches. Visual components like as maps and graphs are developed, along with the relevant labels. Next, visual parameters such as colour, contrast, distance and size are utilised in order to provide a suitable visual hierarchy and a visual path through the information.

Interactivity is becoming an increasingly important component of information visualisation, particularly when it is used in a website or application. The fact that it is interactive enables users to manipulate the visualisation, which makes it exceptionally successful in meeting the requirements of the users. Users are able to see issues from a variety of angles and change their own representations of these topics until they reach the appropriate level of knowledge while using interactive information visualisation. This is especially helpful for those that seek an experience that allows for exploration.



## Notes

### Importance of Information Visualisation

#### 1. Analyzing the Data in a Better Way

Reports, when analysed, enable business stakeholders concentrate their attention on the areas of the company that require it. Visual representations make it easier for analysts to grasp the fundamental concepts relevant to their work. Whether it is a report on sales or a plan for marketing, a visual representation of data assists businesses in increasing their profitability via improved analysis and better business decisions. This is true whether the report is on sales or on marketing.

#### 2. Faster Decision Making

Visual information is easier for humans to process than cumbersome tabular formats or reports. If the data can communicate effectively, decision-makers will be able to move swiftly based on the new data insights, which will speed up decision-making while simultaneously boosting the growth of businesses.

#### 3. Making Sense of Complicated Data

Users of business software can benefit from data visualisation by gaining insight into the large volumes of data they have access to. The ability to discover new patterns and flaws in the data is beneficial to them. By gaining an understanding of these patterns, users are better able to focus their attention on regions that point to warning signs or advancement. This process, in turn, propels the company forward in its goals.

#### 4. Data Visualization Discovers the Trends in Data

The identification of patterns and trends in the data is the primary contribution that data visualisation makes. When all of the data is presented in front of you in a visual manner, as opposed to data that is written out in a table, it is much simpler to recognise patterns and trends within the data.

#### 5. Data Visualization Provides a Perspective on the Data

Data visualisation offers a fresh viewpoint on the information at hand by illuminating its significance within the context of the bigger picture. It illustrates the position of certain data references in relation to the broader image painted by the data.

#### 6. Data Visualization Puts the Data into the Correct Context

While using data visualisation, it is quite challenging to comprehend the data in its natural setting. Seeing statistics in a table by themselves is not enough to have a full understanding of the information since context explains the overall setting in which the data was collected.

### Aesthetics Visualisation

A specific graphic element's aesthetics define every facet of the element itself. Figure 1.1 has a few illustrations to illustrate this point. It should come as no surprise that the location of a graphical element, which specifies where the element may be found, is an essential part of that element. In conventional two-dimensional graphics, locations are denoted by an x and y value; however, it is possible to use a variety of

coordinate systems and to see data in either one or three dimensions. The next thing to note is that every graphical element possesses a form, a size and a colour of its own. Even if we are producing a drawing in black and white, the graphical components still need to have a colour in order to be seen. For example, if the backdrop is white, the graphical elements should be black and if the background is black, they should be white. In conclusion, to the degree that we are utilising lines to show data, these lines may have varying lengths or patterns of dashes and dots depending on the data they represent. In addition to the examples presented in Figure 1.1, we could come across a variety of additional aesthetics while we are examining data visualisations. If we wish to display text, for instance, we could have to define the font family, font face and font size. Similarly, if graphical elements overlap, we might have to declare whether or not they are partially transparent.

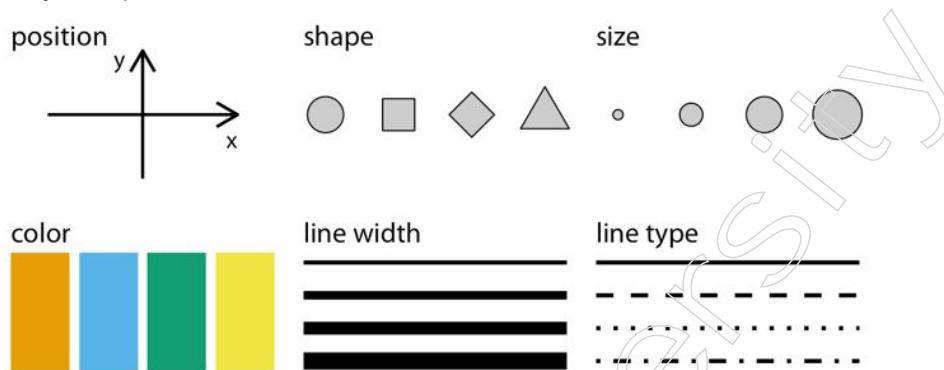


Figure 1.1 demonstrates the common aesthetics that are employed in data visualisation, including location, shape, size, colour, line width and line type. While some of these aesthetics (position, size, line width and colour) are able to convey continuous data as well as discrete data, others of these aesthetics are often only able to represent discrete data (shape, line type).

Any form of aesthetics may be classified as belonging to either of two categories: those that can depict continuous data and those that cannot. Continuous data values are values for which there is the possibility of arbitrarily fine intermediates. For example, time duration is a continuous value. There are an arbitrary number of durations that may be found in the middle of any two durations, such as 50 seconds and 51 seconds. These durations include 50.5 seconds, 50.51 seconds, 50.50001 seconds and so on. On the other hand, the number of people in a room is an example of a discrete variable. It is possible for a room to accommodate either 5 or 6 people, but not 5.5. In the context of the illustrations in Figure 1.1, continuous data may be represented by location, size, colour and line width; however, discrete data can often only be represented by shape and line type.

After this, we will think about the many kinds of data that we would wish to represent in our visualisation. You might think of data as numbers, yet numerical values are only two of the many sorts of data that we could come upon. Data can be presented in a variety of forms, including continuous and discrete numerical values, discrete categories, dates, times and text, in addition to continuous and discrete numerical values (Table 1.2). When information is presented in the form of numbers, we refer to it as quantitative data and when it is presented in the form of categories, we refer to it as qualitative data. Factors are the variables that carry qualitative data, while levels are the numerous categories that may be assigned to factors. Nevertheless, factors can also

## Notes

be sorted when there is an inherent order among the levels of the factor (such as in the example of “good,” “fair,” and “poor” in Table 1.2). In most cases, the levels of a factor do not have an order (as shown in the example of “dog,” “cat,” and “fish” in Table 1.2).

**Table 1.2: Types of variables encountered in typical data visualization scenarios.**

| Type of variable                  | Examples                                     | Appropriate scale      | Description                                                                                                                                                                                                         |
|-----------------------------------|----------------------------------------------|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| quantitative/numerical continuous | 1.3, 5.7, 83, 1.5x10-2                       | continuous             | Arbitrary numerical values. These can be integers, rational numbers, or real numbers.                                                                                                                               |
| quantitative/numerical discrete   | 1, 2, 3, 4                                   | discrete               | Numbers in discrete units. These are most commonly but not necessarily integers. For example, the numbers 0.5, 1.0, 1.5 could also be treated as discrete if intermediate values cannot exist in the given dataset. |
| qualitative/categorical unordered | dog, cat, fish                               | discrete               | Categories without order. These are discrete and unique categories that have no inherent order. These variables are also called factors.                                                                            |
| qualitative/categorical ordered   | good, fair, poor                             | discrete               | Categories with order. These are discrete and unique categories with an order. For example, “fair” always lies between “good” and “poor”. These variables are also called ordered factors.                          |
| date or time                      | Jan. 5 2018, 8:03am                          | continuous or discrete | Specific days and/or times. Also generic dates, such as July 4 or Dec. 25 (without year).                                                                                                                           |
| Text                              | The quick brown fox jumps over the lazy dog. | none, or discrete      | Free-form text. Can be treated as categorical if needed.                                                                                                                                                            |

Have a look at Table 1.3 to get an idea of what each of these different kinds of data looks like in practise. The first few rows of a dataset that provides the daily temperature normals (average daily temperatures over a 30-year span) for four different sites in the United States are displayed here. This table includes five different variables: the month, the day, the location, the station ID and the temperature (in degrees Fahrenheit). Temperature is a continuous numerical value, whereas month is an ordered component, day is a discrete numerical value, location is an unordered factor and station ID is also an unordered factor.

**Table 1.3: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.**

| Month | Day | Location     | Station ID  | Temperature |
|-------|-----|--------------|-------------|-------------|
| Jan   | 1   | Chicago      | USW00014819 | 25.6        |
| Jan   | 1   | San Diego    | USW00093107 | 55.2        |
| Jan   | 1   | Houston      | USW00012918 | 53.9        |
| Jan   | 1   | Death Valley | USC00042319 | 51.0        |
| Jan   | 2   | Chicago      | USW00014819 | 25.5        |
| Jan   | 2   | San Diego    | USW00093107 | 55.3        |
| Jan   | 2   | Houston      | USW00012918 | 53.8        |
| Jan   | 2   | Death Valley | USC00042319 | 51.2        |
| Jan   | 3   | Chicago      | USW00014819 | 25.3        |
| Jan   | 3   | San Diego    | USW00093107 | 55.3        |
| Jan   | 3   | Death Valley | USC00042319 | 51.3        |
| Jan   | 3   | Houston      | USW00012918 | 53.8        |

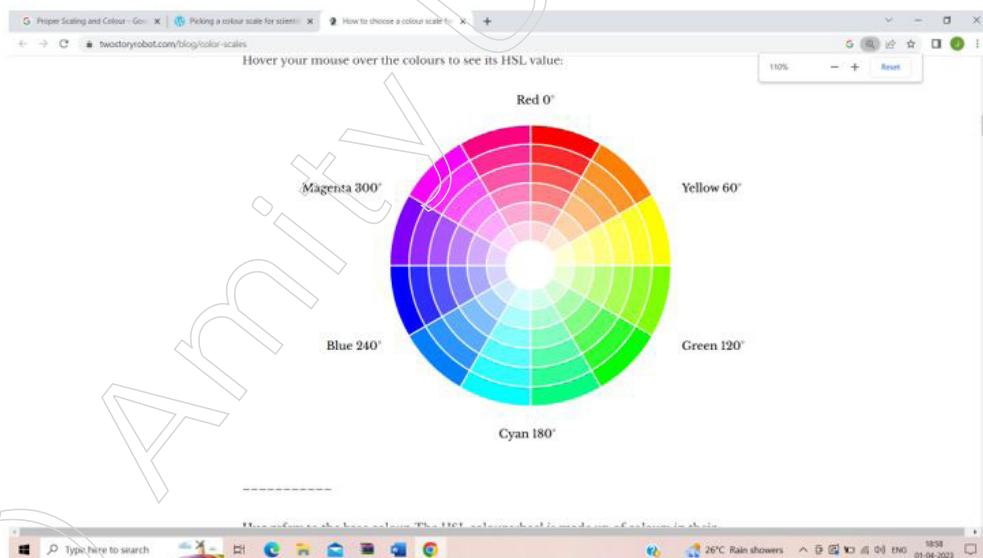
#### 1.4.4 Proper Scaling and Colour, Effective Colour and Shading

##### Understanding HSL

HSL stands for Hue, Saturation and Lightness.

Dealing with colours that are represented by changing degrees of hue, saturation and lightness is a far more natural experience than working with RGB. You should be able to mentally imagine what a colour represented with HSL looks like without having to look at a colour wheel or search the colour up if you have a basic grasp of HSL. This is because HSL is based on hue, saturation and lightness.

Just move your mouse pointer over each colour to get its HSL value:



The term “hue” refers to the primary colour. Because just the fundamental colours are utilised to produce colours in the HSL colour wheel, the colours that make up the wheel are in their most unadulterated form. The creation of these colours did not

## Notes

include the use of any black, white, or grey pigments during the mixing procedure. On the HSL colour wheel, colours are represented by degrees, moving clockwise from red (0 degrees) through yellow (0 degrees), lime (0 degrees), aqua (0 degrees), blue (0 degrees), magenta (0 degrees) and ultimately back to red. Because of this, the hue colour wheel begins with red at 0 degrees and then returns to red at 360 degrees.

The level of the hue's intensity is referred to as the saturation. A colour that has not been diluted by the addition of any of the three primary shades—black, white, or gray—is said to be completely saturated. A fully saturated hue appears on the colour wheel in its purest form, where it possesses the maximum amount of intensity possible. If there is no saturation at all in the colour, then it will seem more like a greyish tone.

The lightness value indicates how brilliant the colour that was selected is. In the same vein as the saturation scale, this too is a percentage scale. Complete darkness, sometimes known as full black, is represented by a brightness value of 0%. Bright white light corresponds to a brightness level of one hundred percent.

This simply indicates that in order to portray a colour in its most unadulterated form, one must choose the hue value of the colour, maintain a saturation of 100% and maintain a lightness of 50%.

### **The three types of colour schemes**

Since we are now familiar with the HSL technique of colour representation, we can investigate the many kinds of colour schemes and learn how to select the appropriate colour scheme for the data that you are attempting to show.

When attempting to visualise data, the use of colour may be a very helpful tool. It is utilised rather frequently in the creation of a wide variety of data visualisations and has the capability of displaying not just correlations and trends but also regions of contrast. It is an interesting medium that, when utilised well, can convey a tremendous lot of information about your data in a way that is both straightforward and easy to understand.

There are many different meteorological and geographical quantities, such as precipitation and political party popularity, that might be beneficial to depict on a map using colour. Each of these quantities, however, requires a distinct sort of colour scheme to be displayed accurately. There are three primary categories of colour schemes: sequential, divergent and qualitative. Each of these colour schemes is best suited for describing a distinct kind of facts. Following that, we are going to investigate these colour schemes by using some interactive visualisations.

#### **a) Sequential colour schemes**

The most effective application of sequential colour schemes is found in quantitative data that can be rationally organised from high to low. When presenting quantitative data, the traditional goal is to demonstrate a progression rather than a contrast between two different values. You are able to display this trend without creating any misunderstanding by making use of a colour scheme that is based on a gradient.

The following graph, which visualised the population size of nations using a sequential colour scheme, may be seen below. The size of the population is one example of the kind of data that works really well with a sequential colouring scheme.

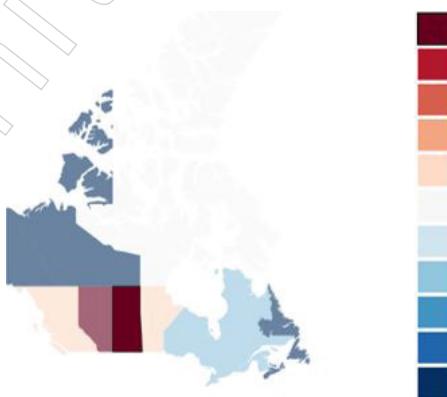


The primary method for producing different colours with a sequential palette is to adjust the level of lightness while maintaining the same hue. In general, brighter colours are related with lighter lightness values, whereas darker colours are associated with greater lightness values.

### b) Diverging colour Schemes

The most effective application of contrasting colour schemes is to bring attention to both high and low extreme values, as well as values that are considerably different from the norm. Likewise, this colour scheme is typically used for data that have a critical middle value (mean, median, or zero value) and a data distribution that includes two ends of relevance. In other words, the data must include a mean, median, or zero value. Different systems place equal focus on extremes at both ends of the data range as well as important values in the middle of the data range.

The following chart provides a graphic representation of the popularity of the Liberal Party and the Conservative Party in the House of Commons in each of Canada's provinces. A colour scheme with contrasting tones is good for this sort of data since the popularity of the two parties might have two extremes that are opposite one another and a middle ground that is neutral. In this scenario, the colour light grey is used to denote the centre ground, while a deep red colour denotes total support for the Conservative party and the opposite is true for the Liberal party.



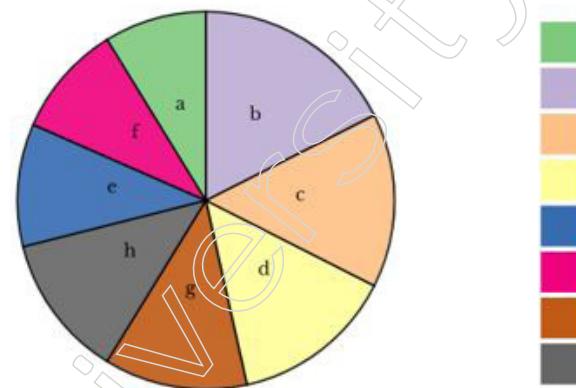
One way to think of a diverging colour scheme is as two successive colour schemes that are oriented in different directions but have their tail ends connected.

## Notes

The point at which they are united from the middle of the diverging colour scheme, which should have a colour that is light and neutral so that darker colours can represent a longer distance from the centre of the diverging colour scheme. It is common practise to assign a unique tone to each of the component sequential palettes. This helps make it simpler to differentiate between values that are positive and values that are negative in relation to the centre.

### c) Qualitative colour Schemes

Since qualitative schemes do not have any intrinsic ordering, they are most effectively utilised for representing nominal or categorical data. It is recommended to limit the number of colours to no more than ten, as having more than that makes it difficult to differentiate between different types. If you have more than ten categories, you might want to think about consolidating some of them into a single category such as "other," as illustrated below.



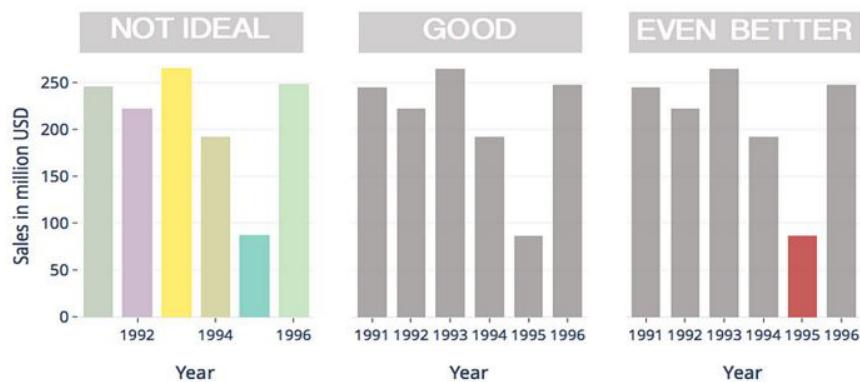
Changing the hues of the colours you choose is the primary method for developing unique colour combinations for a qualitative scheme. In addition, the luminance and saturation of the image can be tweaked ever-so-slightly to further differentiate the colours, however it is recommended that the differences not be made too pronounced. When there is an excessive amount of contrast between the colours, it may give the impression that some colours are more essential than others.

### Rules for optimal use of colour in data visualization: Why colour is key for effective data visualization

The purpose of data visualisation is to facilitate the communication of important findings derived from the process of data analysis. In addition, a chart must have an appealing visual appearance; yet, "looking lovely" is not the primary purpose of a chart. Instead of being a creative endeavour in and of itself, the use of colour in a visualisation should serve the purpose of aiding in the dissemination of essential facts.

#### Rule 1: Use colour when you should, not when you can

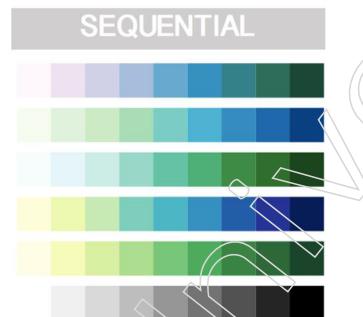
To effectively communicate crucial facts, the use of colour should be properly planned and strategized; hence, this choice cannot be left up to automated algorithms to choose. Bright colours should be reserved for bringing attention to large or uncommon data points, with the majority of the data being shown in neutral hues such as grey.



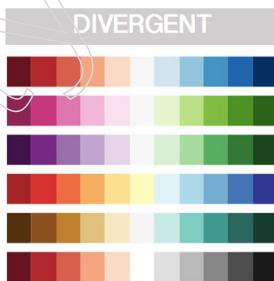
From 1991 through 1996, the sales were recorded in millions of Dollars. The choice of the colour red is intended to attract attention to the exceptionally poor sales that occurred in 1995. All of the almost identical sales from the previous years are depicted in grey.

#### Rule 2: Utilize colour to group related data points

It is possible to use colour to group data points that have a similar value and to depict the extent of this resemblance using the two colour palettes that are listed below:



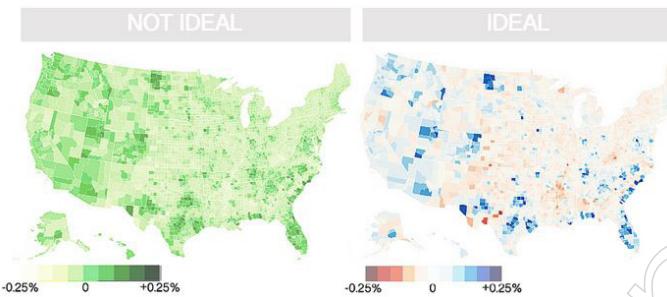
A sequential colour palette is one that consists of increasing intensities of a single hue of colour while maintaining a constant saturation level. The difference in brightness between neighbouring colours is directly proportional to the difference in the data values that those colours are utilised to generate.



A diverging colour palette is created by stacking two successive colour palettes, each of a different hue, one on top of the other and placing an inflection point in the midst of the stack. When trying to visualise data with variances in two opposite directions, they become useful.

The chart on the left utilises a sequential colour palette consisting of a single hue (green) for values ranging from -0.25 to +0.25, whereas the chart on the right employs a divergent colour scheme consisting of separate hues for positive (blue) and negative (red) values. Both charts may be found below.

## Notes



Change, expressed as a percentage, in the population of the United States from 2010 to 2019. The divergent colour scheme, which is composed of two colours (red and blue) and has an inflection point at zero, is superior than the sequential colour scheme in terms of suitability.

While looking at the map on the right, it is possible to tell quickly which values are positive and which are negative simply by looking at the colours. We are able to draw the quick conclusion that the population of towns located in the middle of the country and in the south has decreased, but the population of towns located on the east and west coast has grown. This essential understanding of the data is not immediately apparent in the chart on the left, which requires the reader to focus not on the colour green as such but on the degree to which it is displayed.

### Rule 3: Use categorical colours for unrelated data



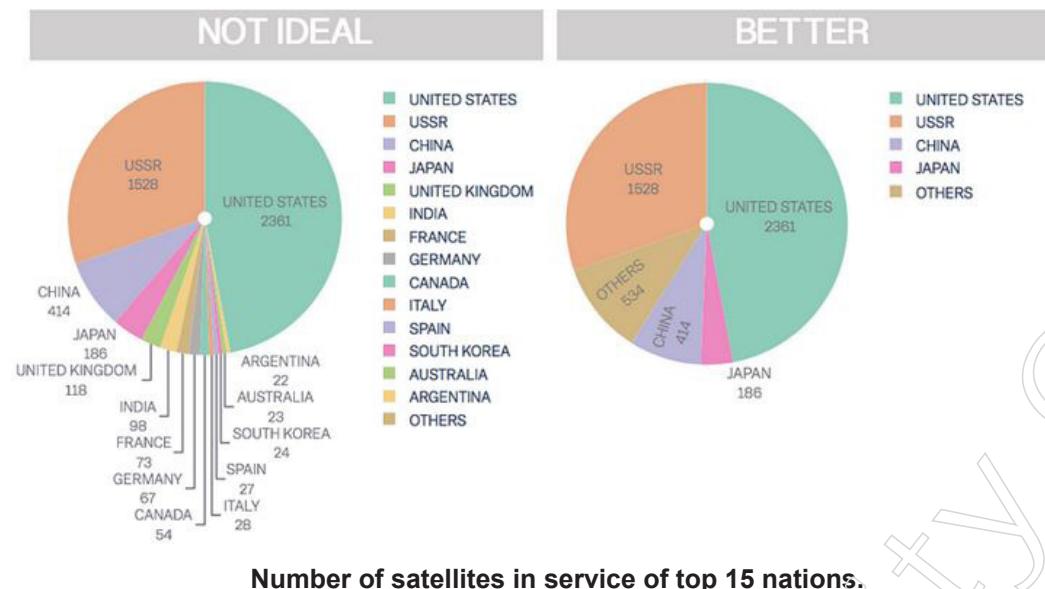
Categorical colour palettes are formed from colours of varied hues but consistent saturation and intensity. These colour palettes may be used to depict data points of entirely different origin or values that are unconnected to one another. Take a look at this graphic representation of the many ethnic groups that may be found in New York City. Since that there is no association between the data pertaining to the various races, a categorical palette has been opted for in this instance.

For displaying variations in magnitude, sequential and divergent colour palettes should be used because they encode qualitative values. On the other hand, categorical colour palettes should be used because they display data categories that are not connected to one another and display quantitative values.

### Rule 4: Categorical colours have few easily discernible bins

Even though utilising a variety of colours might assist in differentiating between various data points, a chart should have no more than six to eight unique colour categories for each of them to be easily recognisable from one another.

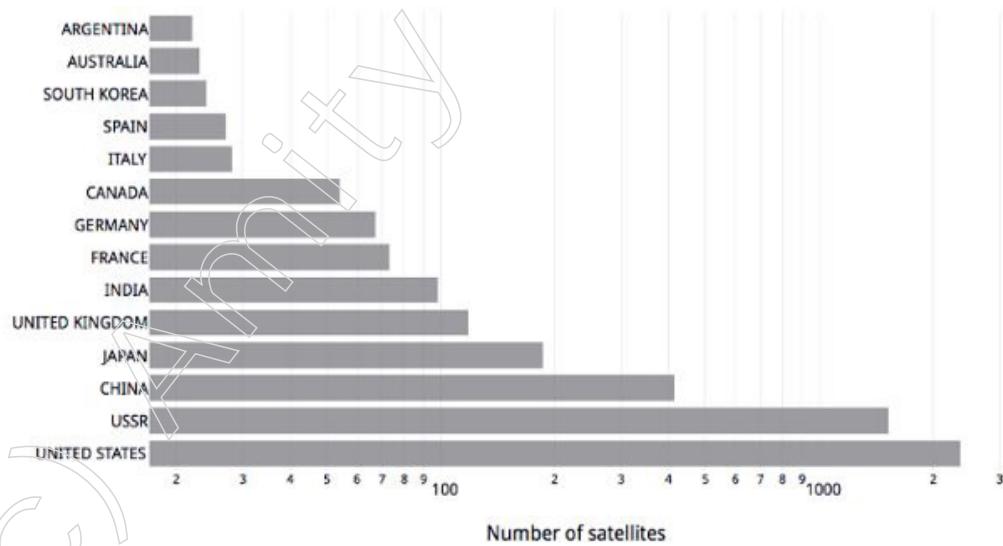
Notes

**Number of satellites in service of top 15 nations.**

The use of a distinct colour for each of the 15 nations makes the chart on the left difficult to understand, particularly with regard to the countries that have a smaller number of satellites. The one on the right is much easier to read, but this comes at the expense of the information on countries who have a smaller number of satellites, which are all put together in the "others" bucket. Please take note that in this instance, we have utilised a categorical colour scheme due to the fact that the data for each nation is totally uncorrelated. For example, the number of satellites that are operated by India is totally separate from those that are operated by France.

#### Rule 5: Change in chart type can often reduce the need for colours

In the example that came before, a pie chart is perhaps not the most appropriate choice. The elimination of categories that follows as a direct consequence is not always appropriate. If we plot the data instead as a bar chart, we may utilise a single colour while still maintaining all 15 data types.

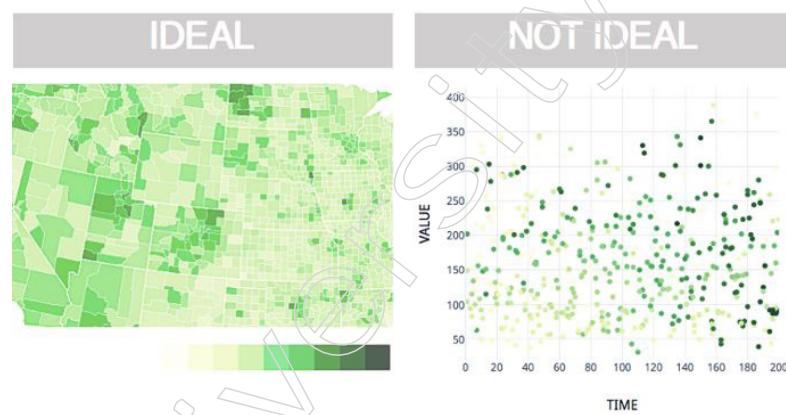
**Number of satellites in service of top 15 nations.**

## Notes

Whenever a visualisation requires more than six to eight distinct colours (hues), either some of the categories should be merged together or more types of charts should be investigated.

### Rule 6: When not to use sequential colour scheme

The colours in a sequential palette need to be arranged such that they are immediately adjacent to one another, just like in the chart on the left of the page below. Only then will the minute differences in colour that occur across the palette become plainly evident. When the data are scattered throughout a plot in a manner similar to a scatter plot, the nuances of the differences become more difficult to understand. The best way to put a sequential colour scheme to use is to display a relative difference in the numbers being displayed. Plotting absolute values, which are more accurately depicted with a categorical colour scheme, cannot be accomplished with this method.



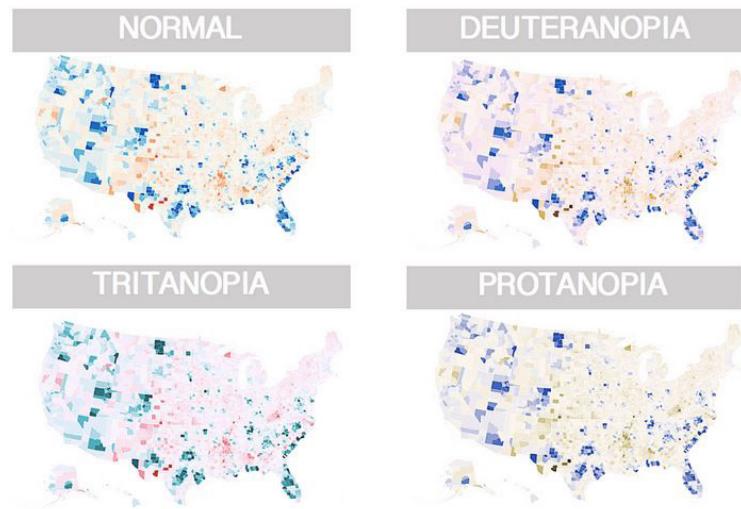
When the data points are not situated exactly next to each other, like they are in the scatter plot on the right, it is difficult to comprehend sequential colour schemes since the colours are sequential. Only for the purpose of visualising relative values, such as in the chart on the left, may these colours be used.

### Rule 7: Choose appropriate background

Check out this animation that was created by Akiyoshi Kitaoka that shows how our perception of the colour of a moving square varies depending on what is happening in the backdrop of the square. The way in which humans see colours is not an absolute. It is determined in relation to the surrounding environment. The colour of an item as seen by the human eye depends not only on the colour of the object itself but also on the colour of the backdrop it is seen against. Because of this, we are forced to reach the following conclusion on the utilisation of backdrop colours in charts. When different things are grouped together by the same colour, the backgrounds of those things ought to be the same. This, in general, indicates that there should be as few differences in the backdrop colour as possible.

### Rule 8: Not everyone can see all colours

It is estimated that around 10% of the population of the globe is colour blind. To ensure that coloured infographics are accessible to everyone, you should avoid using colour combinations that include red and green. What follows is an illustration that shows how three distinct types of colour blindness manifest themselves when viewing the same map.



How colour blindness affects perception of colours.

Notes

## Summary

- The study of data in order to derive useful insights for businesses is referred to as “data science.” To analyse vast volumes of data, this method takes a multidisciplinary approach by combining concepts and methods from the domains of mathematics, statistics, artificial intelligence and computer engineering.
- Apache Spark is a data processing and analytics engine that is open source and can handle massive volumes of data (upwards of several petabytes).
- D3.js is a JavaScript framework that may be used in a web browser to generate individualised representations of data.
- Jupyter Notebook is a web tool that is open source and it enables users to collaborate interactively on projects with other users, including data scientists, data engineers, mathematicians and academics.
- Matplotlib is a plotting library written in Python that is open source and is utilised in analytics applications for reading, importing and visualising data.
- NumPy is an acronym that stands for Numerical Python. It is the name of an open-source Python library that is utilised extensively in applications relating to scientific computing, engineering, data science and machine learning.
- Pandas is yet another well-known open-source Python library. Its primary purpose is to do data manipulation and analysis.
- Python is utilised not just by professionals within the realm of computer, such as data scientists, network engineers and programmers, but also by workers outside of the realm of computing, such as accountants, mathematicians and scientists, who are frequently drawn to its user-friendly character.
- R is a free and open-source platform that may be used for statistical computation and graphical application development.
- The use of data science within businesses makes it possible to analyse and monitor performance criteria, which in turn promotes the development and expansion of the company.

## Notes

### Glossary

- **Descriptive analysis:** The purpose of descriptive analysis is to investigate the data in order to acquire an understanding of what has occurred or what is occurring in the data environment.
- **Predictive analysis:** The goal of predictive analysis is to provide accurate projections about data patterns that may emerge in the future by making use of past data.
- **Data Scrubbing:** The act of standardising the data such that it conforms to a format that has been defined in advance is known as “data cleaning” or “data scrubbing.”
- **Classification:** The process of organising data into distinct groups or categories is known as classification.
- **Regression:** Finding a connection between two data items that at first glance appear to have no bearing on one another is the goal of the statistical technique known as regression.
- **Clustering:** Clustering is a process that involves grouping data that is closely linked together for the purpose of searching for patterns and outliers.
- **Keras:** Keras is a programming interface that simplifies access to and utilisation of the TensorFlow machine learning platform for data scientists. Keras was developed by Google.
- **Matlab:** Matlab is a high-level programming language and analytics environment for numerical computation, mathematical modelling and data visualisation.
- **PyTorch:** PyTorch is a deep learning framework that is open source and is used to develop and train deep learning models that are based on neural networks.

### Check Your Understanding

1. What is the full form of CRM?
  - a) Customer Relationship Management
  - b) Customer Relay Management
  - c) Cluster Relationship Management
  - d) Customer Regression Management
2. What is the full form of DOM?
  - a) Document Object Model
  - b) Different object Model
  - c) Division Object Model
  - d) Data Object Model
3. What is the full form of SPSS?
  - a) System Package for Social Sciences
  - b) System Pytorch for Social Sciences

- c) Statistical Package for Social Sciences  
d) Source Package for Social Sciences
4. What is the full form of SAS?  
a) Sensor Analysis System  
b) Social Analysis System  
c) Statistical Analysis System  
d) Sensor Available System
5. The word \_\_\_\_\_ refers to any type of data that can be saved, retrieved and processed in the form of a predetermined format.  
a) Unstructured data  
b) Structured data  
c) Data Frames  
d) Semi-Structured Data
6. \_\_\_\_\_ refers to any data where the form or structure is unclear. This includes most types of data. Unstructured data presents various obstacles in terms of its processing in order to derive value from it.  
a) Structured data  
b) Optimised Data  
c) Unstructured data  
d) Semi-Structured Data
7. What is the full form of IoT?  
a) Image of Things  
b) Internet of Things  
c) Information of Things  
d) Internet of Times
8. The process of transforming an organisation into a data-driven enterprise is referred to as \_\_\_\_\_.  
a) Cloud Computing  
b) Datafication  
c) Controlling Data Volume  
d) Elastic Compute Cloud
9. The act of analysing the results and drawing conclusions based on data that has been subjected to random fluctuation is known as \_\_\_\_\_.  
a) Simple Storage Service  
b) Sensor Data

**Notes**

**Notes**

- c) Statistical Inference  
d) Social Data
10. \_\_\_\_\_ are defined as a set of values that most likely contains the population value. At this step, the sample error is assessed and a margin is added around the estimate.  
a) Cost Savings  
b) Confidence intervals  
c) Data Science  
d) Social Media Applications
11. \_\_\_\_\_ method use representative samples to evaluate two hypotheses about a population that are incompatible with one another.  
a) Hypothesis Testing  
b) Regression Modelling  
c) Margin of Error  
d) Confidence Intervals
12. \_\_\_\_\_ is a common technique used for conducting objective and accurate analyses of huge data sets. During the examination of enormous volumes of data, the method removes the possibility of error associated with the parameters of the population.  
a) Business Analysis  
b) Probability  
c) Machine Learning  
d) Resampling
13. Sample consists of one or more observations that are drawn from the population and the attribute that may be measured about a sample is referred to as a statistic. The process of choosing a representative sample from a population is referred to as \_\_\_\_\_.  
a) Resampling  
b) Regression  
c) Correlation  
d) Sampling
14. In \_\_\_\_\_, the units of the population being sampled are not chosen at random by the researcher like in other sampling methods.  
a) Point Estimates  
b) Probability Sampling  
c) Confidence Interval Estimates  
d) Estimated Mean

15. \_\_\_\_\_ is an involved process that involves producing sample data and making predictions about the actual world by employing a large number of statistical models and making explicit assumptions.
- a) Statistical modelling
  - b) Pearson Correlation
  - c) Chi-square statistics
  - d) ANOVA or T-test
16. The term \_\_\_\_\_ refers to a strategy that makes use of a single independent variable in conjunction with the best linear correlation to make predictions about a dependant variable.
- a) Multiple Linear Regression
  - b) Bi-variate regression
  - c) Simple Linear Regression
  - d) Multi-variate regression
17. \_\_\_\_\_ is a technique that uses more than one independent variable to make predictions about a dependant variable. This technique provides the best linear connection.
- a) Multiple Linear Regression
  - b) Single Linear Regression
  - c) Lasso Regression
  - d) Logistic Regression
18. The \_\_\_\_\_ is a discrete probability distribution that is used to mimic the frequency of occurrences of an event over a specific length of time or space.
- a) Poisson distribution
  - b) Binomial Distribution
  - c) Normal Distribution
  - d) Fitting a Model
19. The process of portraying data in a meaningful and visually appealing fashion that end users can readily perceive and grasp is referred to as \_\_\_\_\_. This includes graphical representations of data as well as dashboards.
- a) Frequency Distribution
  - b) Variability of Estimates
  - c) Information Visualisation
  - d) Hypothesis Testing
20. The \_\_\_\_\_ of R include sq(), mean() and max(); users may directly call these functions within the application.
- a) Built-in Functions

**Notes**

- b) Functions of variability
- c) Hypothesis Testing
- d) Functions

**Exercise**

1. Explain the concept of Data Science.
2. Explain the concept of Big data.
3. Explain the four V's in Big Data and drivers of Big Data.
4. Explain the role and advantages of Statistics in Data Science.

**Learning Activities**

1. Explain various Data Science Tools.
2. Explain the concept of Statistical Inference.
3. Explain the concept of Population and Samples.

**Check Your Understanding – Answers**

- |        |        |
|--------|--------|
| 1. a)  | 2. a)  |
| 3. c)  | 4. c)  |
| 5. b)  | 6. c)  |
| 7. b)  | 8. b)  |
| 9. c)  | 10. b) |
| 11. a) | 12. d) |
| 13. d) | 14. b) |
| 15. a) | 16. c) |
| 17. a) | 18. a) |
| 19. c) | 20. a) |

**Further Readings and Bibliography**

1. <https://www.jaroeducation.com/blog/three-probability-distribution-used-in-data-science/>
2. <https://h2o.ai/wiki/model-fitting/>
3. <https://www.r-project.org/about.html>
4. <https://www.geeksforgeeks.org/r-programming-language-introduction/>
5. <https://www.geeksforgeeks.org/environments-in-r-programming/>
6. [https://www.tutorialspoint.com/r/r\\_data\\_types.htm](https://www.tutorialspoint.com/r/r_data_types.htm)

## Module - II: Introduction to Data Science

Notes

### Learning Objectives

At the end of this topic, you will be able to understand:

- Analyse descriptive statistics and data preparation
- Learn about exploratory data analysis-summarisation, measuring asymmetry
- Interpret sample and estimated mean, variance, standard score
- Identify statistical inference, frequency approach, variability of estimates
- Analyse hypothesis testing
- Describe chart types: tabular data, dot and line plot, scatter plots, bar plots
- Learn about pie charts, graphs
- Identify description of data using these tools with real time example
- Analyse overview of data science process: defining its goal
- Identify retrieving the data, data preparation-exploration, cleaning and transforming data
- Learn about building the model, presentation and automation
- Analyse introduction and types of machine learning
- Identify role of machine learning in data science
- Interpret classification algorithms: - linear regression, decision tree
- Describe Naive Bayes classifier, K-means
- Analyse K-nearest neighbour, support vector machine

### Introduction

The term “exploratory data analysis,” or EDA, refers to a technique that is utilised by data scientists to evaluate and study data sets as well as summarise the primary characteristics of such data sets. These techniques frequently involve the usage of data visualisation approaches. It makes it simpler for data scientists to see patterns, recognise anomalies, test a hypothesis, or check assumptions by assisting them in determining the most effective way to alter data sources in order to obtain the answers they want.

### 2.1 Philosophy of Exploratory Data Analysis - The Data Science Process

EDA is used largely to examine what the data may disclose beyond the formal modelling or hypothesis testing work and it also gives a deeper knowledge of the variables contained inside a data collection as well as the relationships between those variables. In addition to this, it can assist in determining whether or not the statistical methods that are being considered for the data analysis are suitable. EDA approaches, which were first established as a tool for the process of data discovery in the 1970s

## Notes

by the American mathematician John Tukey, continue to be a method that is frequently utilised today.

### What is Data Science?

If we are able to alter the data in order to uncover previously concealed patterns, it has the potential to be a highly lucrative resource for us. Data Science refers to either the rationale that is behind the data or the technique that lies behind the modification of the data. The process of data science begins with the formulation of a problem statement and continues with the collection of data and the extraction of the required results from that data. The professional who is responsible for determining whether or not this process is proceeding smoothly is known as a Data Scientist. Yet there are also a variety of different career opportunities in this industry, including the following:

1. Data Engineers
2. Data Analysts
3. Data Architect
4. Machine Learning Engineer
5. Deep Learning Engineer

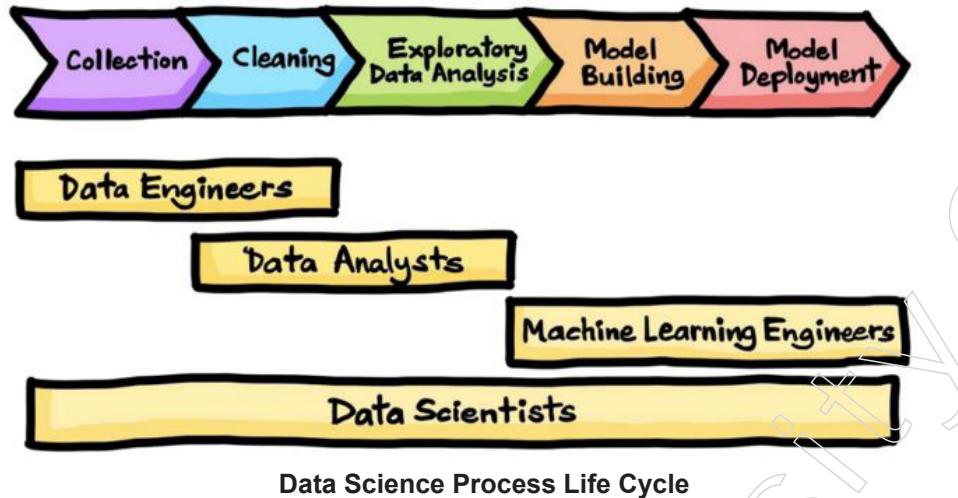
### Data Science Process Life Cycle

There are certain procedures that have to be carried out in order for any of the activities that are being carried out in the subject of data science to produce any successful results from the data that is already available.

**Data Collecting** - Once a problem statement has been formulated, the first activity that has to be completed is to calculate data that will assist us in our analysis and manipulation. Both scraping and surveys are methods that may be used to collect data. Sometimes data is obtained through surveys, while other times it is collected by scraping.

- **Data Cleansing** – Because the vast majority of the data that comes from the actual world is unstructured, it must first be cleaned up and then converted into structured data before it can be utilised for any kind of analysis or modelling.
- **Exploratory Data Analysis** - During this stage of the process, we look for previously unseen patterns in the data we have available to us. In addition to this, we make an effort to do research on the many factors that influence the target variable and the degree to which it does so. All these questions, including how the many independent aspects are connected to one another and what steps might be taken to realise the outcomes that are wanted, can have their answers gleaned from the process. This not only helps us get started with the process of modelling, but it also points us in the direction in which we should work.
- **Model Building** - Several distinct kinds of machine learning algorithms and methods have been developed. They can readily recognise intricate patterns in the data, which is a highly laborious effort for a human to complete and has been made easier with the development of these algorithms and methods.
- **Model Deployment**— Once a model has been constructed and shown to produce

superior outcomes on either the holdout dataset or the real-world dataset, we then deploy the model and monitor how well it is doing. This is the most important step, in which we apply what we have learned from the data to applications and use cases that are based in the actual world.



### Components of Data Science Process

Data science is a very broad field and in order to derive the most value from the data at hand, an individual will need to apply a variety of methodologies and make use of a variety of tools. This will ensure that the data's integrity is maintained throughout the process, while also keeping data privacy in mind. When it comes to machine learning and data analysis, the main focus is on the conclusions that can be drawn from the information that is already available. But, data engineering is the phase of the process in which the primary responsibility is to guarantee that the data is managed appropriately and that appropriate data pipelines are developed to ensure the continuous flow of data. If we were to attempt to list the most important aspects of data science, they would be as follows:

- **Data Analysis** – There are situations in which there is no requirement to apply advanced deep learning and other complicated algorithms to the data that is already available in order to deduce certain patterns from it. Because of this, before we go on to the modelling stage, we first undertake an exploratory data analysis in order to acquire a basic concept of the data and patterns that are accessible in it. This provides us with a direction to work on if we want to apply more complicated analytic methods to our data.
- **Statistics** – The occurrence of a normal distribution in a substantial number of real-world datasets is a natural phenomena. And when we already know that a certain dataset follows some known distribution, then we are able to study the majority of that dataset's attributes at the same time. In addition, descriptive statistics, as well as correlation and covariance analyses between two aspects of the dataset, can assist us in gaining a deeper comprehension of the manner in which one element of our dataset is connected to the other.
- **Data Engineering** - When working with a significant quantity of data, not only do we need to ensure that the data is protected from any potential online risks, but we also need to ensure that it is simple to get the data and to make modifications to

## Notes

it. Data Engineers are an essential component in the process of ensuring that the data are utilised effectively.

- **Advanced Computing**

**Machine Learning** – Machine Learning has opened up new horizons, which has helped us to build a variety of highly advanced applications and methodologies. These advancements have made it possible for machines to become more effective, offer a more individualised experience to each person and perform tasks in the blink of an eye that previously required a great deal of manual labour and time-intensive effort.

**Deep Learning** – This is also a part of Artificial Intelligence and Machine Learning, but it is a bit more sophisticated than machine learning itself. Deep learning refers to the process of learning how to learn. This subfield of data science came into existence as a result of advances in computer power as well as the accumulation of vast amounts of data.

### 2.1.1 Descriptive Statistics and Data Preparation

Descriptive statistics are statistics that explain, demonstrate and summarise the fundamental characteristics of a dataset that may be discovered in particular research. These statistics are provided in a summary that summarises the data sample and its measurements. The analysts are able to better interpret the data as a result. Statistics that are only descriptive provide a representation of the data sample that is currently available; they do not include any hypotheses, judgements, probabilities, or conclusions. Inferential statistics are the right tool for the task here.

#### Descriptive Statistics Examples

The grade point average of a student is a great example of descriptive statistics and you don't have to search any farther to find one (GPA). The cumulative grade point average (GPA) is a metric that is used to evaluate a student's overall academic performance since it takes into account all of the grades, classes and tests that the student has taken and calculates an average out of those numbers. Take into consideration that the GPA does not give any conclusions or even attempt to forecast future performance. In its place, it offers a simple overview of the academic progress of pupils based on values drawn from the data.

Here is a case that is considerably easier to understand. Let's say that the total of a data set consisting of 2, 3, 4, 5 and 6 is equal to 20. The number four was determined to be the set's mean by dividing the total by the total number of values (20 divided by 5 equals 4).

While presenting descriptive statistics, analysts frequently make use of charts and graphs. Descriptive statistics are the type of data you would get if you went outside of a movie theatre, polled fifty people about how much they like the movie they just saw and then plotted the results on a pie chart. In this particular scenario, descriptive statistics count the number of "yes" and "no" responses to determine the percentage of viewers in this particular theatre who enjoyed or did not enjoy the movie. If you tried to arrive to any other conclusions, you would be wandering into the domain of inferential statistics; however, we will address that topic later on in this discussion.

In conclusion, political polling is regarded a descriptive statistic as long as it just presents tangible facts (the responses supplied by the respondents) and does not make any judgements based on the findings. The following is an explanation of how polls work: "Who did you choose to be the next President in the election that just took place?"

### Types of Descriptive Statistics

Statistics that are descriptive can be broken down into several different sorts, features, or measurements. There are two sorts, according to the claims of certain writers. Some people say three, while others claim even four.

### Distribution (Also Called Frequency Distribution)

A data set is made up of a collection of scores or values in a certain format. The frequency of each conceivable value of a variable is summarised by statisticians using graphs and tables, which can display the data as percentages or raw figures. For instance, if you were to conduct a survey to find out which Beatle fans prefer, you would first build up two columns: the first would include all of the available variables (John, Paul, George and Ringo) and the second would provide the total number of votes.

### Measures of Central Tendency

The average or centre of a dataset may be estimated using measures of central tendency and the outcome can be found using one of three methods: the mean, the mode, or the median.

The mean, which is commonly referred to as "M," is the approach that is utilised the vast majority of the time for determining averages. The mean is calculated by first adding together all of the response values and then dividing that total by the total number of replies, denoted by the letter "N." Consider the following scenario: someone is trying to calculate how many hours they sleep each day during the course of a week. Hence, the entries for the hours (such as 6,8,7,10,8,4,9) would make up the data set and the total of those values would be 52. As there are seven replies, we may deduce that N equals 7. To get M, which is 7 in this case, you take the value sum of 52 and divide it by N, which is 7.

Mode: The mode is basically the most frequent answer value. There is no limit to the amount of modes that a dataset may contain, even "zero." You may determine the mode of your dataset by first sorting the values in descending order from the lowest to the highest and then looking for the answer that occurs the most frequently. Thus, utilising the results of our study on sleep from the previous section: 4,6,7,8,8,9,10. As can be seen, the most common value is eight.

Median: At long last, we've arrived at the median, which is the number that is exactly in the middle of the whole dataset. Get the number that is in the middle of the set by sorting the values such that they increase in value from lowest to highest (just like we did with the mode). In this particular instance, the median is 8.

### Variability (Also Called Dispersion)

The statistician is able to get an indication of the degree to which the responses are dispersed thanks to the measure of variability. The range, the standard deviation and the variance are the three components that make up the spread.

## Notes

You can figure out how far away the most extreme numbers are by using range, which measures the distance between two values. To begin, take the range of values in the dataset and remove the lowest value from the greatest value. Once more, let's look at the results of our sleep study: 4, 6, 7, 8, 9, 10. We obtain the number six by taking the lowest, which is four and subtracting it from the greatest, which is 10. That is the scope of your ability.

Departure from the norm: A little bit more effort is required for this component. The standard deviation, denoted by the letter "s," represents the average amount of variability in your dataset. It demonstrates the distance each score is from the overall average. The bigger the standard deviation of your data, the more varied your group of numbers will be. Take the following six steps:

1. Compile a list of the scores and their respective meanings.
2. Determine the standard deviation by taking each score and subtracting the mean from it.
3. Do a square root of each deviance.
4. Compute the sum of the squared deviations for each value.
5. Divide the total squared deviations by  $N-1$  and then take the quotient.
6. Determine the square root of the result.

### Univariate Descriptive Statistics

Univariate descriptive statistics focus on analysing just one variable at a time and do not make any comparisons between the variables. Instead, it gives the researcher the opportunity to characterise the factors individually. As a consequence of this, the statistics of this kind are sometimes referred to as descriptive statistics. The following can be used to provide an explanation for the patterns that have been found in this kind of data:

- Values that represent the average of a group (mean, mode and median)
- Data dispersion (standard deviation, variance, range, minimum, maximum and quartiles) (standard deviation, variance, range, minimum, maximum and quartiles)
- Bar graphs
- Pie graphs
- Frequency polygon histograms
- Tables of frequency distribution

### Bivariate Descriptive Statistics

While utilising bivariate descriptive statistics, two variables are examined (compared) simultaneously to determine whether or not they are connected. The columns are often used to represent the independent variable, while the rows are used to represent the dependent variable, as this is the standard convention.

Bivariate data may be utilised in a wide variety of contexts in the real world. For instance, it might be highly useful to make an educated guess about when a natural event would take place. A statistician's toolkit should include bivariate data analysis

as one of the options. There are instances when something as straightforward as projecting one parameter against the other on a two-dimensional plane might help you better comprehend what the information is attempting to persuade you of. For instance, the following scatterplot illustrates the correlation between the amount of time that elapses between eruptions at Old Faithful and the total amount of time that an eruption lasts.

### 2.1.2 Exploratory Data Analysis-Summarisation, Measuring Asymmetry

John Tukey is credited for popularising exploratory data analysis as a means of motivating statisticians to investigate existing data and maybe develop hypotheses that could lead to the conduct of additional research and experiments. EDA has a more focused focus, with an emphasis on validating the assumptions that are necessary for model fitting and testing hypotheses. In addition, it verifies while it is dealing with missing data and transforming variables according to the requirements. EDA creates a comprehensive knowledge of the data as well as any problems connected to either the information or the process. This method takes a scientific approach to figuring out what the facts are trying to tell us.

#### Types of Exploratory Data Analysis:

1. Univariate Non-graphical
  2. Multivariate Non-graphical
  3. Univariate graphical
  4. Multivariate graphical
1. **Univariate Non-graphical:** This is the simplest type of data analysis since during this type of study, we just look at one variable at a time when researching the data. The usual objective of univariate non-graphical EDA is to get an understanding of the underlying sample distribution and data in order to then draw conclusions about the population. The identification of outliers is an extra component of the study. The following are some of the features of population distribution:
- **Central tendency:** The central tendency, also known as the place of distribution, relates to values that are typical or in the centre of the range. The most frequent and helpful measurements of central tendency are statistics called mean, median and sometimes mode, with mean being the most common of the three. Sometimes mode can also be useful. When there are concerns about outliers or when the distribution is skewed, the median may be the most appropriate measure to use.
  - **Spread:** The spread is a measure of how far out from the centre we are in our search for the information values. Spread may be thought of as a percentage. The standard deviation of the quality as well as the variance are two relevant measurements of dispersion. The variance is equal to the mean square of the individual deviations and as a result, the variance is equal to the variance's root.
  - **Skewness and kurtosis:** The skewness and kurtosis of the distribution are two additional univariates descriptors that may be quite helpful. As compared

## Notes

to a normal distribution, skewness is the measure of asymmetry, while kurtosis may be a more nuanced indicator of how peaked the distribution is.

2. **Multivariate Non-graphical:** A multivariate non-graphical EDA approach is one that is often used to demonstrate the link between two or more variables using cross-tabulation or statistics. This technique is referred to as “multivariate non-graphical” EDA.
  - An extension of tabulation known as cross-tabulation is an exceptionally valuable tool for analysing categorical data. Cross-tabulation is the method of choice when there are two variables involved. This method involves creating a two-way table with column headings that correspond to the amount of one variable and row headings that correspond to the amount of the other two variables. Then, the counts are filled in with all the subjects that share an equivalent pair of levels.
  - We construct statistics for quantitative variables independently for every level of the particular variable, then compare those statistics across the amount of categorical variables after creating the statistics for quantitative variables for each category variable and one quantitative variable.
  - Comparing the means is an impromptu method of performing an ANOVA, although comparing the medians may be an accurate method of performing a one-way ANOVA.
3. **Univariate graphical:** Although non-graphical methods are quantitative and objective, they are unable to provide a comprehensive picture of the data. As a result, graphical methods are used more frequently because they require a degree of subjectivity in their analysis and are therefore preferred. Non-graphical methods are quantitative and objective. The following are examples of common types of univariate graphics:
  - **Histogram:** A histogram is the most fundamental type of graph. A histogram is a type of bar plot in which each bar reflects either the frequency (count) or the percentage (count divided by total count) of cases for a range of values. Histograms are one of the easiest and quickest ways to understand a significant amount about your data, including its central tendency, spread, modality, form and outliers.
  - **Stem-and-leaf plots:** stem-and-leaf plots are a simple alternative that can be used in place of a histogram. It reveals all the data values and, consequently, the form of the distribution.
  - **Boxplots:** The boxplot is an additional highly helpful univariate graphical approach. Boxplots are great for providing information about symmetry and outliers, as well as presenting robust measures of location and spread. However, they can be misleading regarding aspects such as multimodality. Boxplots are excellent at providing information about central tendency and show robust measures of location and spread. The usage of boxplots in the form of side-by-side boxplots is one of the most straightforward applications of boxplots.
  - **Quantile-normal plots:** These are the most complex univariate graphical EDA approach available. They are used to analyse data based on quantiles and normal distributions. It is known as the quantile-normal plot, abbreviated

as QN, or more commonly as the quantile-quantile plot, abbreviated as QQ. It is common practise to evaluate how well a particular sample conforms to a certain theoretical distribution. It makes it possible to identify deviations from normal as well as diagnose skewness and kurtosis.

4. Multivariate graphical: Multivariate graphical data makes use of visuals in order to demonstrate correlations between two or more kinds of information. The only one that is typically used is a grouped bar plot, in which each group represents one level of one of the variables and each bar inside a group represents the quantity of the other variable. This type of bar chart is the most popular.

Additional typical examples of multivariate graphics include the following:

- **Scatterplot:** The primary graphical EDA tool for two quantitative variables is the scatterplot. This technique thus has one variable on the x-axis and one on the y-axis and consequently the point for every example in your dataset.
- A run chart is a line graph that displays data drawn over a period of time.
- A heat map is a graphical representation of data in which values are shown by colour. It is also known as a temperature map.
- The multivariate chart is a graphical depiction of the connections between the many factors and the responses to those factors.
- Bubble charts are a type of data visualisation that depict many circles, or bubbles, in a two-dimensional space.

In a nutshell: Before continuing with any additional analysis of your data, you should always carry out the proper EDA. Take whatever measures are necessary to become more familiar with your data, check for errors that are readily apparent, educate yourself on the distributions of the variables and educate yourself on the connections between the variables. EDA is not a perfectly accurate science; yet it is vitally significant.

### Tools Required for Exploratory Data Analysis

The following is a list of some of the most often used tools in the process of creating an EDA:

1. **R:** R is a free and open-source programming language that is used for statistical computation and graphics. R is backed by the R foundation for statistical computing. In order to construct statistical observations and conduct data analysis, statisticians frequently make use of the R programming language.
2. **Python:** Python is an object-oriented programming language that can be interpreted and has dynamic semantics. Its high level, built-in data structures, in conjunction with dynamic binding, make it a particularly appealing option for the quick creation of applications. Moreover, it may be used as a scripting or glue language to tie together pre-existing components due to its versatility. Python and EDA are frequently used in conjunction with one another to identify missing values in a data collection. This is an essential step before you can determine how to deal with missing values for machine learning.

In addition to these functions, which have already been detailed, EDA is also able to:

## Notes

- Carry out k-means clustering: k-means clustering is an unsupervised learning process in which the information points are allocated to clusters. It is also known as k-groups and it is typically used in market segmentation, picture compression and pattern recognition.
- EDA is frequently used in predictive models, such as linear regression, where it is used to predict results. EDA may be used to predict outcomes.
- In univariate, bivariate and multivariate visualisation, it is also applied for summary statistics, the establishment of linkages between each variable and the comprehension of how various fields within the data interact with one another.

### Measures of Skewness

The degree to which the individual values deviate from the mean is reflected in the asymmetry measure. In a symmetrical distribution, the items exhibit a perfect balance on either side of the mode, while in a skew distribution, the balance is thrown to one side. In contrast, a normal distribution exhibits a perfect balance on both sides of the mode. The degree to which the sum of the two sides is greater than the balance is used to quantify the skewness of the series. One easy approach to describe skewness in a series is to use the difference between the mean, the median and the mode. In the event that the skewness is positive, we get  $Z < M < X$  and in the event that the skewness is negative, we have  $X < M < Z$ . In most cases, this is how we assess the skewness:

$$\text{Skewness} = X - Z \text{ and its coefficient } (j) \text{ is worked}$$

In case  $Z$  is not well defined, then we work out skewness as under:

$$\text{Skewness} = 3(X - M) \text{ and its coefficient } (j) \text{ is worked}$$

$$\begin{aligned} \text{Skewness} &= \bar{X} - Z \text{ and its coefficient } (j) \text{ is worked} \\ \text{out as } j &= \frac{\bar{X} - Z}{\sigma} \end{aligned}$$

In case  $Z$  is not well defined, then we work out skewness as under:

$$\text{Skewness} = 3(\bar{X} - M) \text{ and its coefficient } (j) \text{ is worked}$$

$$\text{out as } j = \frac{3(\bar{X} - M)}{\sigma}$$

When the elements of a particular series are plotted on a graph, the significance of skewness lies in the fact that through it, one can study the formation of series and can have an idea about the shape of the curve, whether it is normal or not. Skewness also allows one to study the formation of series. The degree to which a curve has a flat top is measured using kurtosis. If a curve is substantially more peaked than the normal curve, then we name it Leptokurtic and if a curve is relatively flatter than the normal curve, then we call it Platykurtic. A bell-shaped curve or the normal curve is mesokurtic since it is kurtic in the centre. Kurtosis, or the humpedness of the curve, is a statistical term that describes the manner in which the items that fall in the centre of a series are distributed.

Since the majority of techniques make certain assumptions about the nature of the distribution curve, it is important to be aware of the shape of the distribution curve before applying statistical approaches to the analysis of research data. This is because of the fact that the majority of methods.

### 2.1.3 Sample and Estimated Mean, Variance, Standard Score

- **Sample and Estimated Mean**

The mean of a group of data is referred to as a sample mean. Calculating the central tendency, standard deviation and variance of a data set are all possible using the sample mean as the starting point. The sample mean may be utilised for a number of purposes, including the estimation of population averages, among other applications. A wide variety of professional fields make use of statistical data as well, including the following:

- Areas of study in the scientific world like as biology, ecology and meteorology
- All aspects of medicine and pharmacy
- Computer and data science, information technology and computer and network security
- Industries related to space travel and aviation
- Fields in engineering and design

The sample mean is a measurement that indicates where the centre of the data lies. The sample mean is used to produce an approximation of the mean of any population. We are obliged to make an estimate of what the entire population is doing, or what all of the components going across the population, in a number of scenarios and cases, while not being able to conduct a survey with each individual member of the population. The sample mean can be helpful in situations like these. The word "sample mean" refers to the average value that may be obtained in a sample. Having determined the sample mean, the next step is to compute the variance and from there, the standard deviation.

$$\text{Sample Mean} = (\text{Sum of terms}) \div (\text{Number of Terms})$$

$$\begin{aligned} &= \frac{\sum x_i}{n} \\ &= \frac{(x_1 + x_2 + x_3 + \dots + x_n)}{n} \end{aligned}$$

#### How to calculate the sample mean

The sample mean may be easily calculated by first counting the number of items in a sample set, then adding those numbers together and then dividing that total by the total number of items in the sample set. You may use the following formula to determine the sample mean using whatever spreadsheet software or calculator you prefer:

$$\bar{x} = (\Sigma x_i) / n$$

Here,  $\bar{x}$  represents the sample mean,  $\Sigma$  tells us to add,  $x_i$  refers to all the X-values and  $n$  stands for the number of items in the data set.

In order to calculate the sample mean utilising the formula, you will need to enter in the values that correspond to each of the symbols. The calculation of the sample mean of a data collection may be broken down into the following phases for your reference:

##### 1. Add up the sample items

To begin, you will need to determine how many sample items are contained inside a data set and then sum up the total number of things that are contained within the set. Consider the following illustration:

## Notes

Consider the following scenario: a professor is interested in determining the typical grade attained by his students. The example set provided by the instructor has seven possible test scores, which are as follows: 78, 89, 93, 95, 88, 78 and 95. After tallying up all of the points, he arrives with the total of 616. In the subsequent phase, which is the determination of the sample mean, he may make use of this total.

### 2. Divide sum by the number of samples

After that, divide the total from the previous step by the overall quantity of the items in the data collection. Following the example of the instructor, here is what this looks like in practise:

To calculate the class average, for instance, the instructor adds up all 616 possible points. Because there were seven total scores in his data set, he divides 616 by seven to get the answer. The quotient that was arrived at is 88.

### 3. The result is the mean

Following division, the quotient that is obtained is the sample mean, often known as the average. Take, for instance, the case of the educator: For instance, the student's scores, which he was calculating at the time, resulted in an average grade of 88 percent. The sample mean can be used as a starting point for additional calculations of the variance, standard deviation and standard error.

### 4. Use the mean to find the variance

You may utilise the sample mean in additional calculations by first determining the sample's own variance, then using the sample mean in those calculations. The term "variance" refers to the degree to which each of the sample items in a data collection are dispersed from one another. Finding the difference between each data item and the mean is the first step in calculating the variance of the data. Let us use the example of the instructor to illustrate how this works:

Example: The instructor wants to determine the range of his students' scores, so he begins by calculating the difference between the average score and each of the seven students' scores that he used to get the mean:  $(78-88, 89-88, 93-88, 95-88, 88-88, 78-88 \text{ and } 95-88) = (-10, 1, 5, 7, 0, -10, 7)$ .

After that, the instructor squares each difference ( $100, 1, 25, 49, 0, 100, 49$ ), puts all the numbers together and then divides the total by seven in the same manner as the mean. When he divides 324 by 7, he gets 46.3, which is almost the same as 46. The greater the variance, the greater the degree to which the data deviates from its mean.

### 5. Use the variance to find the standard deviation

In addition, you have the option of computing the standard deviation of the sample set to take the sample mean one step further. The square root of the variance is the standard deviation and it is used to describe the rate at which a collection of data follows the normal distribution. Consider the following illustration:

Example: To calculate the standard deviation, the instructor utilises the variance of 46, which equals 6.78. This figure indicates to the instructor how far above or below the class average of 88% the student in question is on any specific test result that is included in the sample set.

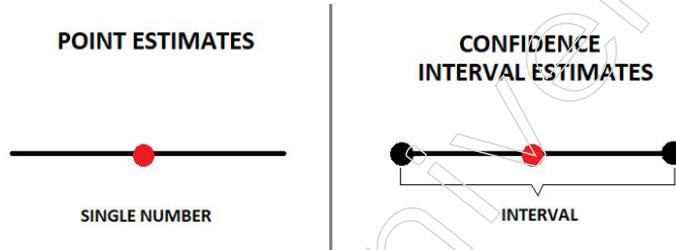
### Estimated Mean

The process of drawing conclusions about a population parameter based on information obtained from a sample is referred to as estimation. A point estimate is considered to be the most accurate estimate, despite the fact that it is derived from a sample of a population that is chosen at random. In addition, if you repeatedly collect random samples from the same population, you should anticipate that the point estimate will change from sample to sample. This is because it is reasonable to assume that the population is not constant.

A confidence interval, on the other hand, is an estimate that is formed on the premise that the real parameter will fall within a given proportion regardless of the number of samples that are analysed. This assumption is made regardless of the number of samples that are analysed. An estimate is a particular value, but a population estimator is an approximation that is based purely on sample information. On the other hand, a population estimator is referred to as an estimate.

There are two distinct kinds of estimations:

- Point Estimates— single number.
- Confidence Interval Estimates — Provide much more information and are preferred when making inferences.



As the point estimate is located in the middle of the confidence interval estimate, there is a connection between the two. On the other hand, confidence intervals offer a great deal more information and are the method of choice for drawing conclusions.

- Variance

The term variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean (average) and thus from every other number in the set. Variance is often depicted by this symbol:  $\sigma^2$ . It is used by both analysts and traders to determine volatility and market security.

The standard deviation (SD or  $\sigma$ ), also known as the square root of the variance, is a statistic that may be used to assess how stable an investment's returns have been over a certain amount of time. In statistics, variability is measured by comparing individual values to an average or mean. To compute it, first the differences between each number in the data set and the mean are taken, then the differences are squared so that they have a positive value and lastly the sum of the squares is divided by the total number of values in the data set.

## Notes

Variance is calculated by using the following formula:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{N}$$

**where:**

$x_i$  = Each value in the data set

$\bar{x}$  = Mean of all values in the data set

$N$  = Number of values in the data set

### Steps for Calculating the Variance

Regardless of the type of software you employ for your statistical study, the variance is often computed mechanically on your behalf. But, if you want to get a better grasp of how the formula works, you may also calculate it by hand. The process of determining the variance manually consists of five primary phases. In order to guide us through the process, we will utilise a tiny data set consisting of only six scores.

| Data set |    |    |    |    |    |
|----------|----|----|----|----|----|
| 46       | 69 | 32 | 60 | 52 | 41 |

**Step 1:** Finding the mean is the first step.

The mean may be determined by first adding together all of the scores and then dividing that total by the total number of points.

**Mean ( $\bar{x}$ )**

$$\bar{x} = (46 + 69 + 32 + 60 + 52 + 41) \div 6 = 50$$

**Step 2:** Determine how far each score is above or below the mean.

To calculate the deviations from the mean, take the mean score and subtract it from each individual score.

Since  $\bar{x} = 50$ , deduct 50 from each person's score.

| Score | Deviation from the mean |
|-------|-------------------------|
| 46    | $46 - 50 = -4$          |
| 69    | $69 - 50 = 19$          |
| 32    | $32 - 50 = -18$         |
| 60    | $60 - 50 = 10$          |
| 52    | $52 - 50 = 2$           |
| 41    | $41 - 50 = -9$          |

**Step 3:** Square each departure from the mean for the third step.

Do a separate multiplication for each value that is different from the mean. This will result in numbers that are more than zero.

### Squared deviations from the mean

$$(-4)^2 = 4 \times 4 = 16$$

$$19^2 = 19 \times 19 = 361$$

$$(-18)^2 = -18 \times -18 = 324$$

$$10^2 = 10 \times 10 = 100$$

$$2^2 = 2 \times 2 = 4$$

$$(-9)^2 = -9 \times -9 = 81$$

**Step 4:** Determine the sum of the squares in the fourth step.

The total squared deviations should be added together. The name for this concept is the “sum of squares.”

### Sum of squares

$$16 + 361 + 324 + 100 + 4 + 81 = 886$$

**Step 5:** Divide the total number of squares by  $n - 1$ , often known as N

Divide the total number of squares by  $n - 1$  (if you’re doing this for a sample) or by N. (for a population).

Since we are working with a sample, we will utilise  $n - 1$ , where  $n = 6$ .

### Variance

$$886 / (6 - 1) = 886 / 5 = 177.2$$

- **Standard Score**

A datum or an observation, the standard score makes up a portion of the standard deviation and is itself a score. Both the score and the standard deviation may be easily compared due to the fact that they share numerous commonalities. The depiction of the standard score in a manner that is favourable will be the datum that is located in the upper position of the mean. When looking at the mean, placing the datum lower on the scale implies that the score is lower than it should be.

### What is a Standard Score?

The standard score is used to calculate the distance between the score corresponding to the standard deviation and the unit mean. There are several other names for the score, including standardised variables, Z score, normal score and a few more. The difference between the population mean and the raw score is what the Z score attempts to measure. Both the populations and the raw score are represented by the standard deviations, each in its own unit. The intelligence test is the best illustration to use to comprehend the standard score. The mean is a unit of the score and its value is the overall average of all the unit scores. The score contributes to the definition of the specific number that exists between the mean, denoted by ( $M$  or  $\mu$ ) and the score of standard deviations, denoted by ( $s$  or  $\sigma$ ). To compute the standard score, you will need the mean and standard deviation of the raw score.

### Standard Score: Characteristics

The standard score provides access to a large number of significant qualities, such as the fact that it may be utilised in the process of score distribution for comparison purposes. The distribution technique will produce the mean, the standard deviations, as well as the raw materials. The majority of the time, the score is compared to other

## Notes

scores in an accurate manner. The second aspect of the score is that larger deviations are generated from a higher standard score. This is the case for both positive and negative aspects of the score. The unit mean is comprised of the additional raw score in addition to the larger standard score.

The fact that the score went from one to zero indicates that it remained on the mean unit throughout. The difference between a score and its mean, as measured by standard deviations, is referred to as the standard score. This is another characteristic's definition. If the z score is positive, then the score will be greater than the mean even if the score itself is in a negative condition. The score can also be in a positive state. During the point in time when the Z score is in a negative state, the score will be lower than the Mean.

It is not possible to establish the importance of the Z score just on whether the indications are positive or negative. An example, +1.0 is smaller than the -2.0 Z score.  $Z=+1$  has a smaller distance with the unit mean but  $Z = -2.0$  has a double distance with the unit mean. The negative or positive signs of the Z score determine the score distance from the unit mean. The exact value of the Z score determines the magnitude of the standard score.

### Standard Score: Formula

The formula for determining the standard score is as follows:  $Z=X/\sigma$ , which is given below:

$$Z=X-\mu/\sigma$$

The formula that was just presented explained that the score ( $X$ ) is subtracted from the unit mean and then the difference between the two is divided. The letter Z stands for the standard score and the computation is done using the raw score. Normal distribution is used. One category has a standard deviation of  $S = 1$ , whereas another category has a standard deviation of  $S = 5$ . Both the  $S = 1$  based class, which will receive a score of  $X = 80$  and the  $S = 5$  based class will receive a score of  $X = 85$ . The mean or average score for both courses will be seventy-five, while the mean score for the second class will be eighty. The equation reads as follows:  $Z = 80 - 75/1 = 5$ , whereas the standard score for another class reads as follows:  $Z = 85 - 80/5 = 1$ . The score can be used to compare the overall results of two different classes. At the point in time when the score is being distributed, the score ( $X$ ) is translated into the scoring standard, which is denoted by Z.

### Applications of Standard Score

In order to calculate standard score, one must first have an understanding of the connection between standard deviation and the unit mean. The many uses of the scores are explained in the following paragraphs.

- The score is the most often used standardised approach and it plays a role in helping to make pupils comparable to one another. The work of determining the total population might, at times, appear to be rather challenging; however, with the assistance of a standard score, the task can be simplified.
- There are two intervals of prediction that go into calculating the Standard Score.

These intervals are the lower endpoint and the higher endpoint. The observation of the population that will exist in the future provides an indicator of these two periods.

- The off-target operation of the method is controlled by the process constant.

### Standard Score: Advantages

Using the standard score comes with a number of benefits that will be discussed below:

- The score contributes to the process of determining the value of the raw data based on the unit mean and the standard deviation unit. Consider the possibility that a standard score of two indicates that the value of the standard deviation is likewise two.
- When comparing two sets of data, the score is a very helpful tool to have on hand. The score is utilised in the computation of the relative value as well as the likelihood within the normal standard distribution.

### Standard Score: Disadvantages

The following list describes a few of the drawbacks associated with utilising a standard score:

- The standard score is not capable of distinguishing between ordinal and nominal forms of data.
- There is no way for the score to reconstruct the data's initial values. Standard deviations and distributions can be used to assist in the process of recalculating the values.

## 2.1.4 Statistical Inference, Frequency Approach, Variability of Estimates

- Statistical Inference

The act of analysing the results and drawing conclusions based on data that has been subjected to random fluctuation is known as statistical inference. Inferential statistics is another name for this method. The applications of statistical inference include the testing of hypotheses and the calculation of confidence intervals. The process of drawing conclusions about the characteristics of a population based on the results of a random sample is known as statistical inference. It is helpful in evaluating the link between the variables that are dependent and those that are independent. It is the goal of statistical inference to arrive at an estimate of the uncertainty or the variance from sample to sample. Because of this, we are able to produce a probable range of values for the actual levels of anything that is prevalent in the population. The following criteria are taken into consideration when drawing conclusions based on statistical data:

- Sample Size
- Variability in the sample
- Size of the observed differences

## Notes

### Types of Statistical Inference

There are many distinct kinds of statistical inferences and many of them are utilised in the process of coming to conclusions. They are as follows:

- One sample hypothesis testing
- Confidence Interval
- Pearson Correlation
- Bi-variate regression
- Multi-variate regression
- Chi-square statistics and contingency table
- ANOVA or T-test

### Statistical Inference Procedure

The following are the steps that are involved in inferential statistics:

- You should start with a theory.
- Formulate a working hypothesis for the research
- Ensure that the variables are operationalized.
- Identify the group of people to whom the findings of the study should be applicable.
- Provide a testable alternative to the null hypothesis for this group.
- Collect a representative sample of the population, then carry on with the research.
- Carry out statistical tests to determine whether or not the attributes of the gathered samples are sufficiently distinct from those that would be anticipated on the basis of the null hypothesis in order to be able to reject the null hypothesis.

### Statistical Inference Solution

Statistical inference methods result in the effective utilisation of statistical data linked to populations of persons or experiments. It covers all of the characters, as well as the gathering, examination and analysis of data, as well as the organisation of the data that has been gathered. After beginning work in a variety of sectors, individuals are able to acquire information via the use of statistical inference solutions. Some facts regarding statistical inference solutions include the following:

- It is usual practise to make the assumption that the observed sample is comprised of independent observations drawn from a population type such as Poisson or normal.
- The statistical inference solution is utilised in order to determine the value(s) of the parameter(s) of the anticipated model, such as the normal mean or the binomial proportion.

### Importance of Statistical Inference

Inferential The correct examination of the data requires an understanding of statistics. Interpreting the findings of the research requires careful data analysis to ensure an appropriate conclusion can be drawn from it. Its primary use is in the forecasting of future events for a wide range of data in a variety of domains. It facilitates

the process of drawing conclusions based on the facts. The statistical inference has a wide range of applications in a variety of fields, including the following:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector

- **Frequency Approach**

There are several examples of frequency distribution in our everyday lives. Nearly every profession, including the meteorological department, data scientists and civil engineers, makes use of frequency distributions in their work. Because of these distributions, we are able to draw conclusions from any data set, identify prevailing patterns and forecast upcoming values as well as the general trajectory of the data. There are two varieties of frequency distributions: grouped and ungrouped. Each have their advantages and disadvantages. Its application is contingent on the data with which we are currently working. The examination of their findings is an extremely vital component of both probability and statistics. Let us look at each of these ideas in more depth.

### Frequency Distributions

The distribution of frequencies over the values may be understood through the use of frequency distributions. It is the number of values that are contained within each of the intervals. They provide us with an idea of the range in which the majority of the values lie as well as the ranges in which there are few values. A frequency distribution is a summary of all the possible values of a variable together with the frequency with which each value occurs.

There are a few different types of frequency distributions:

1. **Grouped Frequency Distributions:** In this method, values are first segmented into a variety of intervals and then the frequencies of each segment are tallied.
2. **Un-Grouped Frequency Distributions:** This type of frequency distribution lists each unique value of the variable and counts the frequency with which it occurs.

**Question:** Consider the following scenario: we have statistics on the number of goals scored by a team over 10 separate matches.

1, 0, 0, 3, 2, 0, 2, 3, 1, 1

Create a frequency table to serve as a visual representation of this facts.

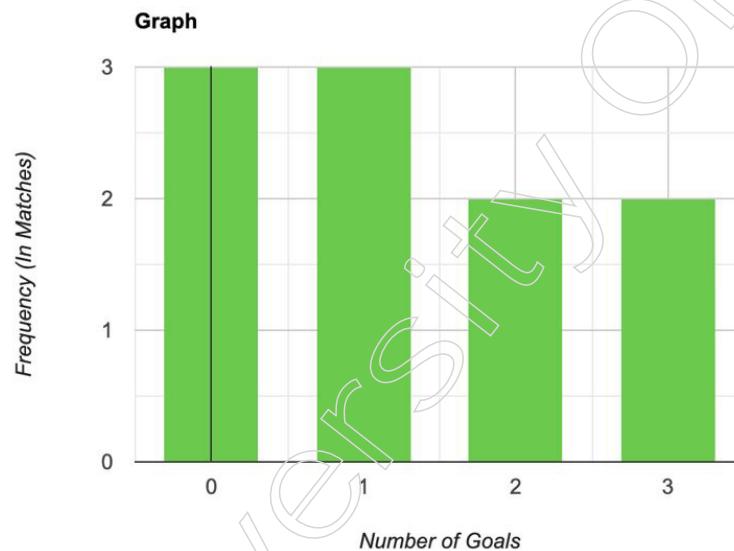
**Solution:**

Since there are a smaller number of values that are distinct. It is not necessary for us to group the data. Just counting the unique values and the frequency with which they occur is sufficient.

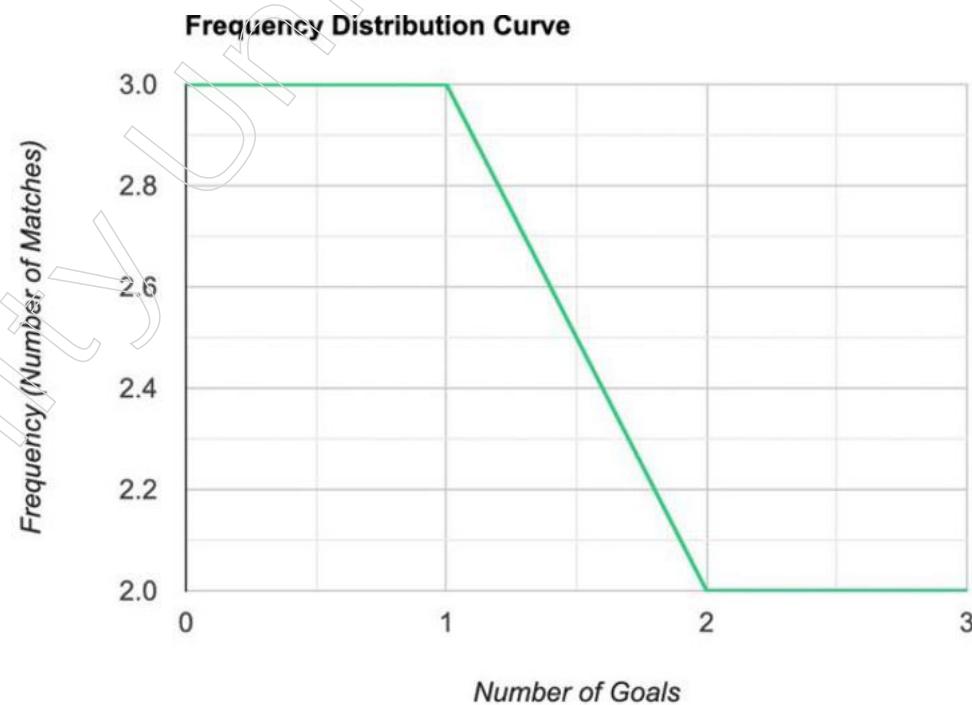
## Notes

| Number of Goals | Frequency |
|-----------------|-----------|
| 0               | 3         |
| 1               | 3         |
| 2               | 2         |
| 3               | 2         |
| Total           | 10        |

This frequency table can also be represented in the form of a bar graph.



A frequency distribution can also be represented by a line curve. The figure given below represents the line curve for the above problem.



In a similar vein, if there are a large number of unique values, we can classify them into groups and generate grouped frequency distributions much like we did in the prior scenario.

## Cumulative Frequency Distribution

The definition of cumulative frequency is the total of all the frequencies that have occurred in the values or intervals that have come before the present one. The frequency distributions that are represented by the cumulative frequencies are termed cumulative frequency distributions. This is because cumulative frequencies are used to depict frequency distributions. The cumulative frequency distribution may be broken down into two distinct categories:

1. Less than type: we add up all of the frequencies that occurred before the present interval.
2. More than type: we add together all of the frequencies that occurred after the most recent interval.

Let us have a look at an example to discover how to properly express a cumulative frequency distribution.

Question 1: The values of the runs that Virat Kohli has scored in the last 25 T-20 matches are listed in the table below. The data should be presented in the form of a cumulative frequency distribution of the less than type:

|    |    |    |    |    |
|----|----|----|----|----|
| 45 | 34 | 50 | 75 | 22 |
| 56 | 63 | 70 | 49 | 33 |
| 0  | 8  | 14 | 39 | 86 |
| 92 | 88 | 70 | 56 | 50 |
| 57 | 45 | 42 | 12 | 39 |

### Solution:

When there are many different values, the solution is for us to describe this information in the form of grouped distributions with intervals such as 0-10, 10-20 and so on. Initially, let's try to make sense of the data by presenting it in the form of a clustered frequency distribution.

| Runs   | Frequency |
|--------|-----------|
| 0-10   | 2         |
| 10-20  | 2         |
| 20-30  | 1         |
| 30-40  | 4         |
| 40-50  | 4         |
| 50-60  | 5         |
| 60-70  | 1         |
| 70-80  | 2         |
| 80-90  | 2         |
| 90-100 | 1         |

Now that we have a frequency distribution, we will turn it into a cumulative frequency distribution by adding the values of the current interval to those of all the intervals that came before it.

## Notes

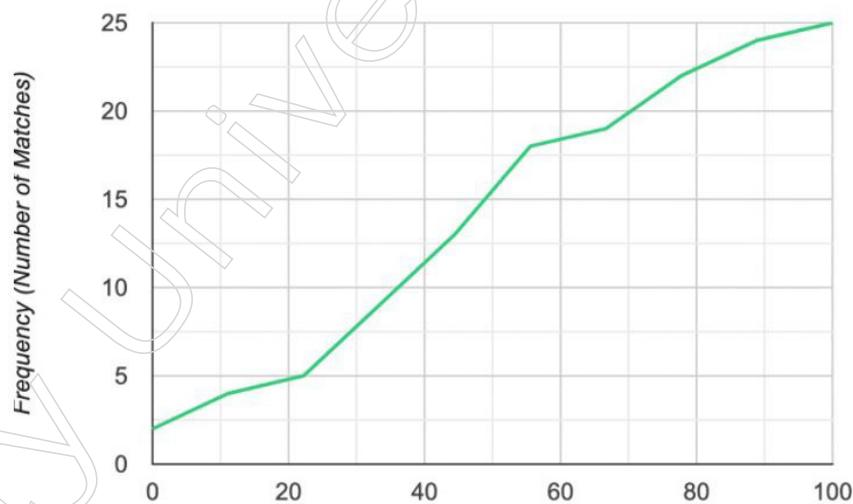
| Runs   | Frequency |
|--------|-----------|
| 0-10   | 2         |
| 10-20  | 4         |
| 20-30  | 5         |
| 30-40  | 9         |
| 40-50  | 13        |
| 50-60  | 18        |
| 60-70  | 19        |
| 70-80  | 21        |
| 80-90  | 23        |
| 90-100 | 25        |

The cumulative frequency distribution is represented here in this table.

**Question 2:** Convert the cumulative frequency distribution table that was just shown into the form of a line curve that represents the cumulative frequency distribution.

**Solution:** The solution is to utilise the point in the middle of each interval as well as the value associated with it when plotting the line curve for the table above.

**Cumulative Frequency Distribution Curve**



- **Variability of Estimates**

In the field of statistics, the term “variability” refers to the divergence of scores in a group or series from their respective mean values. It is more accurate to say that it relates to the variance of the group’s scores in comparison to the mean. Another name for this phenomenon is dispersion. For instance, in a group of ten individuals who have all received different grades on a mathematics test, there are differences between each person in terms of the total number of points that he or she has obtained. These changes may be assessed with the assistance of a measure of variability, which measures the dispersion of distinct values for the average value or average score. Measures of variability quantify the spread of values around the mean. The dispersion of the values within a group is another definition of variability or dispersion. A high degree of variability in the distribution indicates that the scores are not consistent with one another and are spread out over a large range. When there is little variability, it

indicates that the scores are comparable and consistent with one another, as well as centred in the middle.

According to Minium, King and Bear's (2001) research, measures of variability provide a quantitative representation of the degree to which the scores in a distribution either cluster together or disperse around. They do not define the distance that a specific score deviates from the group's average, but rather they reflect the dispersion of a whole collection of results. The shape of a distribution or the degree of performance of a group cannot be determined using these measures of variability since they do not supply such information. The branch of statistics known as descriptive statistics encompasses measures of variability. These statistics reflect the degree to which a collection of scores is comparable to one another.

If the scores were more comparable to one another, the measure of variability, also known as dispersion, would be reduced accordingly. The measure of variability or dispersion will be larger if there is less similarity between the scores than there is between themselves. In general, the measure of dispersion will be bigger if there is a distribution that is more spread out. To put it clearly, dispersion is the variance that exists between the data values that are contained inside a sample. The range and the standard deviation are the two measurements of dispersion that are utilised the vast majority of the time. In the previous lesson, we spoke about different ways to measure central tendency.

Yet, despite the fact that measurements of central tendencies are quite important, their applications are rather restricted. Even if we may compare two or more groups by using these measures, the comparison of two or more groups requires more than just a measure of central tendency. They do not display the manner in which the individual scores are distributed. Let us have a look at another illustration that is comparable to the one that was covered in the section titled "Introduction." A teacher of mathematics is curious about the levels of achievement attained by two groups (A and B) of his or her pupils. They are given an exam that is worth forty points. The following is a breakdown of the test scores attained by the students in groups A and B respectively:

Marks of Group A: 5,4,38,38,20,36,17,19,18,5 (N = 10, Total = 200, Mean = 20)

Marks of Group B: 22,18,19,21,20,23,17,20,18,22 (N = 10, Total = 200, Mean = 20)

There is no significant difference in the overall performance of the two groups, as indicated by the fact that the mean score for both sets of results is 20. Yet, there is a distinction between the performance of the two groups in terms of the degree to which the marks earned by each individual student differ from those earned by the other student. For example, the range of test results for group A was found to be anywhere from 5 to 38, whereas the range for group B was found to be anything from 18 to 23. It indicates that some of the students in group A are performing exceptionally well, others are performing extremely poorly and yet others are performing at a level that is approaching the level of the average student.

On the other hand, the performance of all of the students who are part of the second group is falling inside and near approximately the average (mean) value of 20, which is 20. This demonstrates that the measurements of central tendency only provide us an imperfect view of the data set that we are looking at. It provides an inadequate foundation on which to construct a comparison of two or more sets of scores. We thus

## Notes

require, in addition to a metric for determining the central tendency of the data, an index that indicates the degree to which the scores are dispersed around the mean of the distribution.

In other words, we require a method for calculating the dispersion or variability of the data. A summary of the scores is what constitutes a measure of central tendency, whereas a description of the scores' dispersion is what constitutes a measure of dispersion. It is frequently just as vital to have information about the variability as it is to have information about the core trend. Because we are concerned with the arithmetic mean of the deviations from the mean of the values of the individual items, the phrase "variability" or "dispersion" is also known as the "average of the second degree."

This is due to the fact that we take into consideration these deviations here. So, in order to accurately characterise a distribution, we often need to offer a measure of both the central tendency and the variability of the distribution. In the process of statistical inference, measures of variability are quite significant. The variability in random sampling may be understood better with the use of several measurements of dispersion. When taking a random sample, how much variation should be expected? This question, which concerns the subject's inherent variability, is essential to the resolution of each and every issue pertaining to statistical inference.

The following are some of the reasons why it is necessary to have metrics of variability:

- The extent to which the properties of a data set are represented by an average may be evaluated with the use of various measures of variability. If there is just a little amount of variance, this suggests that the values in the distribution are quite consistent and the average will accurately represent the properties of the data. On the other hand, if the variance is high, this suggests that there is a reduced degree of consistency and that the average is not accurate.
- Variability measures are helpful in determining the nature of variation as well as the factors that contribute to it. This kind of information can be beneficial in helping to control the fluctuation.
- Variability measures are helpful in comparing the spread of two or more data sets with regard to how uniform or consistent the data sets are.
- The use of additional statistical methods such correlation, regression analysis and so on is made easier by the utilisation of measures of variability.

### Functions of Variability

The following is a list of the primary roles that dispersion or variability serves:

- It is utilised in the process of computing other types of statistics, including analysis of variance, degree of correlation, regression and a variety of others.
- It may also be used to compare the variability of the data that was acquired, such as in the case of Socio-Economic Status, income, education and so on.
- To determine whether the calculated average as well as the mean, median and mode are accurate. If the variance is low, then we can assert that the average that was computed is accurate; but, if the variation is high, then it is possible that the average was calculated incorrectly.

- The dispersion helps us regulate the variability by providing us with an indication of whether the data are being negatively impacted by the variability.

### 2.1.5 Hypothesis Testing

The statistical method known as “Hypothesis Testing” involves putting your presumptions about a population parameter to the test in order to determine whether or not they are accurate. It is utilised in the process of estimating the degree to which a link exists between 2 statistical variables.

Let us take a look at a few real-world instances of statistical hypotheses, shall we?

- A professor at his college speculates that sixty percent of the students at his institution originate from households that fall under the lower middle class.
- According to one medical professional, the diabetes treatment known as “3D” (diet, dosage and discipline) has a success rate of 90 percent.

#### How Hypothesis Testing Works

An analyst puts a statistical sample through a series of tests as part of the hypothesis testing process. The purpose of these tests is to provide evidence on the plausibility of the null hypothesis. Statistical analysts put a theory to the test by measuring and analysing a representative sample drawn at random from the population under investigation. The “null hypothesis” and the “alternative hypothesis” are the two hypotheses that are tested by every analyst using a population sample that is chosen at random.

A hypothesis of equivalence between population parameters is an example of a null hypothesis; for instance, a null hypothesis would claim that the population mean return is equal to zero. An alternative hypothesis can be thought of as the theoretical antithesis of a null hypothesis (e.g., the population mean return is not equal to zero). As a result, the two cannot be true at the same time since they contradict one another. Nonetheless, one of the two theories will invariably turn out to be accurate.

#### 4 Steps of Hypothesis Testing

The following are the steps that are taken to test each hypothesis:

- The first thing that the analyst will do is express the two hypotheses in a way that allows just one of them to be correct.
- The next stage is to construct an analysis strategy, which will describe how the data will be analysed and what conclusions will be drawn from them.
- The third stage is to put the strategy into action and conduct a thorough examination of the sample data.
- The fourth and last stage is to conduct an analysis of the results and either conclude that the null hypothesis cannot be supported by the data or declare that the null hypothesis is consistent with the evidence.

#### Topic 2.2 Basic Tools of EDA (Plots, Graphs and Summary Statistics)

##### Introduction

## Notes

Exploratory data analysis, often known as EDA, is one of the methods utilised in the field of data science with the purpose of identifying key characteristics and trends that are then employed by machine learning and deep learning models. As a result, EDA has developed into a significant benchmark for everyone working in the field of data science.

### The Significance of EDA in the Field of Data Science

The discipline of data science is currently highly essential in the world of business since it offers numerous options to make crucial business choices by evaluating large amounts of obtained data. This is why the field has become so important. For a complete understanding of the data, it is necessary to investigate it from every angle. EDA holds a priceless position in the field of data science because of its impacting qualities, which enable users to make decisions that are meaningful and helpful.

### Objective of Exploratory Data Analysis

In most cases, the sub-goals of exploratory data analysis are broken down into the following categories since the primary goal of exploratory data analysis is to get crucial insights.

- Locating and eliminating any anomalies in the data.
- Seeing patterns across both time and space.
- Locate patterns that are associated with the goal.
- Forming theories and putting those hypotheses to the test via various experiments.
- Locating brand new information gathering sources.

### Role of EDA in Data Science

The utilisation of the aforementioned objectives forms the basis for the function that data exploration and analysis play. When the data have been formatted, the analysis that is conducted reveals patterns and trends that are helpful in determining the appropriate actions that are necessary to fulfil the anticipated goals of the organisation. In the same way that we expect specific responsibilities to be completed by every executive working in a given job role, we also anticipate that appropriate EDA will provide comprehensive responses to questions concerning a specific business decision. As constructing models for prediction is an integral part of data science, it is necessary for those models to take into account the most relevant aspects of the data. As a result, EDA guarantees that the appropriate components, in the form of patterns and trends, are made accessible for training the model in order to obtain the desired result, analogous to the way a good recipe works. As a result, realising the desired outcome will be easier if the appropriate EDA is carried out using the appropriate tool and is based on data that is appropriate.

### Steps Involved in Exploratory Data Analysis (EDA)

The primary activities that constitute an EDA's performance are referred to as its essential components. The following describe each of these:

## 1. Data Collection

In today's world, data is being produced in vast quantities and in a wide variety of formats and this occurs in every aspect of human existence, including medicine, athletics, industry, tourism and so on. Every company understands the need of utilising data in a productive manner by effectively evaluating it. Nevertheless, this is contingent on the successful collection of the necessary data from a variety of sources, including but not limited to customer evaluations, social media and surveys. It is not possible to move on with more activities unless adequate and pertinent data are collected.

The housing dataset is represented here by the data that can be found below. It provides details about properties, including the prices at which they were sold.

|      | <b>Id</b> | <b>MSSubClass</b> | <b>MSZoning</b> | <b>LotFrontage</b> | <b>LotArea</b> | <b>Street</b> | <b>Alley</b> | <b>LotShape</b> | <b>LandContour</b> | <b>Utilities</b> | <b>... </b> | <b>PoolArea</b> | <b>PoolQC</b> | <b>Fence</b> | <b>MiscFeature</b> | <b>MiscVal</b> |
|------|-----------|-------------------|-----------------|--------------------|----------------|---------------|--------------|-----------------|--------------------|------------------|-------------|-----------------|---------------|--------------|--------------------|----------------|
| 0    | 1         | 60                | R.L.            | 65.0               | 8450           | Pave          | NaN          | Reg             | lsl                | AllPub           | ...         | 0               | NaN           | NaN          | NaN                | 0              |
| 1    | 2         | 20                | R.L.            | 80.0               | 9600           | Pave          | NaN          | Reg             | lsl                | AllPub           | ...         | 0               | NaN           | NaN          | NaN                | 0              |
| 2    | 3         | 60                | R.L.            | 60.0               | 11250          | Pave          | NaN          | IR1             | lsl                | AllPub           | ...         | 0               | NaN           | NaN          | NaN                | 0              |
| 3    | 4         | 70                | R.L.            | 60.0               | 9550           | Pave          | NaN          | IR1             | lsl                | AllPub           | ...         | 0               | NaN           | NaN          | NaN                | 0              |
| 4    | 5         | 60                | R.L.            | 84.0               | 14260          | Pave          | NaN          | IR1             | lsl                | AllPub           | ...         | 0               | NaN           | NaN          | NaN                | 0              |
| ...  | ...       | ...               | ...             | ...                | ...            | ...           | ...          | ...             | ...                | ...              | ...         | ...             | ...           | ...          | ...                | ...            |
| 1455 | 1456      | 60                | R.L.            | 62.0               | 7917           | Pave          | NaN          | Reg             | lsl                | AllPub           | ...         | 0               | NaN           | NaN          | NaN                | 0              |
| 1456 | 1457      | 20                | R.L.            | 85.0               | 13175          | Pave          | NaN          | Reg             | lsl                | AllPub           | ...         | 0               | NaN           | MinInv       | NaN                | 0              |
| 1457 | 1458      | 70                | R.L.            | 65.0               | 9042           | Pave          | NaN          | Reg             | lsl                | AllPub           | ...         | 0               | Wall_GarPv    | Shed         | 2500               |                |
| 1458 | 1459      | 20                | R.L.            | 65.0               | 9717           | Pave          | NaN          | Reg             | lsl                | AllPub           | ...         | 0               | NaN           | NaN          | NaN                | 0              |
| 1459 | 1460      | 20                | R.L.            | 75.0               | 9937           | Pave          | NaN          | Reg             | lsl                | AllPub           | ...         | 0               | NaN           | NaN          | NaN                | 0              |

1460 rows × 81 columns

**Figure: Housing Dataset**

## 2. Finding all Variables and Understanding Them

The accessible data, which contain a wealth of information, are the primary subject of attention at the outset of the analysis process. This information has varying values for a variety of qualities or attributes, which makes it easier to comprehend them and get insightful information from them. It is necessary to first determine the significant factors that influence the outcome as well as the potential impact those factors may have. This stage is essential to achieving the desired outcome, which may be predicted from any analysis.

## 3. Cleaning the Dataset

The next step is to clean the data set, which may contain null values and other information that is not pertinent to the study. They are to be eliminated so that the data contains just the values that are significant and pertinent to the intended use from the perspective of the target. This will not only cut down on the amount of time required, but it will also cut down on the amount of processing power required. During preprocessing, all problems are resolved, including the identification of null values, outliers and anomaly detection, among other things.

## 4. Identify Correlated Variables

To better understand the relationship between two variables, it is helpful to identify any correlations that may exist between them. The correlation matrix approach provides a distinct depiction of the ways in which various variables are correlated, which assists further in comprehending the critical connections that exist between them.

## Notes

## 5. Choosing the Right Statistical Methods

As you will see in following parts, different statistical tools are utilised based on the data, whether they are categorical or numerical, the size of the data set, the kind of variables and the reason for conducting the analysis. The information that is obtained through applying statistical equations to numerical outputs is valid, but graphical visualisations are more appealing and easier to understand.

## 6. Visualizing and Analyzing Results

When the analysis has been completed, the results need to be scrutinised with extreme caution and attention so that an accurate interpretation may be derived from them. The patterns that emerge in the distribution of the data and the association between the variables provide valuable insights that may be used to make adjustments that are appropriate for the data parameters. The data analyst has to have the necessary analytic capabilities and should be familiar with all of the different approaches to analysis. The findings that were acquired will be suited for the data of that particular field and they are applicable to the fields of retail, healthcare and agriculture.

## 2.2.1 Chart Types: Tabular Data, Dot and Line Plot, Scatter plots, Bar plots, Pie Charts, Graphs

- Tabular Data

The information required for tabulation of data is the kind of information that may be found in spreadsheets and CSV files. In most cases, they are arranged in the form of rows and columns. As contrast to photos or text, this kind of data makes up a significant portion of the datasets from which companies attempt to derive value. Examples of this data include sensor readings, clickstreams, purchasing patterns and databases used for customer management.

In the field of statistics, the term “tabular data” refers to information that is laid down in the form of a table, complete with rows and columns.

## Tabular Data

**columns = attributes for those observations**

The rows of the table reflect the observations, whereas the columns of the table indicate the properties associated with those observations.

Tabular data can be represented, for instance, by the table that follows:

**Tabular Data**

columns = attributes for those observations

| Player | Minutes | Points | Rebounds | Assists |
|--------|---------|--------|----------|---------|
| A      | 41      | 20     | 6        | 5       |
| B      | 30      | 29     | 7        | 6       |
| C      | 22      | 7      | 7        | 2       |
| D      | 26      | 3      | 3        | 9       |
| E      | 20      | 19     | 8        | 0       |
| F      | 9       | 6      | 14       | 14      |
| G      | 14      | 22     | 8        | 3       |
| I      | 22      | 36     | 0        | 9       |
| J      | 34      | 8      | 1        | 3       |

The rows in this dataset total 9 and there are 5 columns.

Each row is an individual basketball player and each of the five columns describes a distinct characteristic of that player. These characteristics are as follows:

- Player name
- Minutes played
- Point
- Rebounds
- Assists

### When is Tabular Data Used in Practice?

Tabular data is the most typical sort of data that you will come across in the real world while you are collecting information. In the real world, the vast majority of the data that is stored in an Excel spreadsheet is regarded to be tabular data. This is due to the fact that the rows in the spreadsheet represent observations, while the columns represent the qualities associated with those observations.

For instance, the following is how our basketball dataset that we discussed previously may appear in a spreadsheet created in Excel:

## Notes

|    | A      | B       | C      | D        | E       | F | G | H |
|----|--------|---------|--------|----------|---------|---|---|---|
| 1  | Player | Minutes | Points | Rebounds | Assists |   |   |   |
| 2  | A      | 41      | 20     | 6        | 5       |   |   |   |
| 3  | B      | 30      | 29     | 7        | 6       |   |   |   |
| 4  | C      | 22      | 7      | 7        | 2       |   |   |   |
| 5  | D      | 26      | 3      | 3        | 9       |   |   |   |
| 6  | E      | 20      | 19     | 8        | 0       |   |   |   |
| 7  | F      | 9       | 6      | 14       | 14      |   |   |   |
| 8  | G      | 14      | 22     | 8        | 3       |   |   |   |
| 9  | I      | 22      | 36     | 0        | 9       |   |   |   |
| 10 | J      | 34      | 8      | 1        | 3       |   |   |   |
| 11 |        |         |        |          |         |   |   |   |
| 12 |        |         |        |          |         |   |   |   |
| 13 |        |         |        |          |         |   |   |   |
| 14 |        |         |        |          |         |   |   |   |
| 15 |        |         |        |          |         |   |   |   |
| 16 |        |         |        |          |         |   |   |   |
| 17 |        |         |        |          |         |   |   |   |
| 18 |        |         |        |          |         |   |   |   |
| 19 |        |         |        |          |         |   |   |   |
| 20 |        |         |        |          |         |   |   |   |
| 21 |        |         |        |          |         |   |   |   |
| 22 |        |         |        |          |         |   |   |   |
| 23 |        |         |        |          |         |   |   |   |
| 24 |        |         |        |          |         |   |   |   |

Since using this format to gather and store values in a dataset is one of the most logical methods to do it, you'll notice that it's used rather frequently.

- **Dot and Line Plot**
  - **Dot Plot**

A dot plot, also known as a strip plot or dot chart, is an easy way to visualise data that consists of data points displayed as dots on a graph that has an x- and y-axis. Another name for this type of data visualisation is a dot chart. This category of charts is utilised to graphically illustrate particular data patterns or groups. The Federal Reserve's quarterly estimates for interest rates are maybe the most well-known dot plot in existence. These projections are provided by the Federal Reserve.<sup>1</sup> A dot plot is like a histogram in that it shows the distribution of a collection of data by displaying the number of data points that fall into each category or value on an axis. This is done in the same way that a histogram does.

### Types of Dot Plots

The Cleveland and Wilkinson dot plots are the two most important forms of dot plots. Both make use of dots, but there are significant distinctions between the two; Wilkinson is more comparable to a histogram, whilst Cleveland is more similar to a bar graph.

### Cleveland Dot Plot

The variable is represented as continuous in the Cleveland dot plot, as opposed to being represented as a categorical variable. This is comparable to a bar chart,

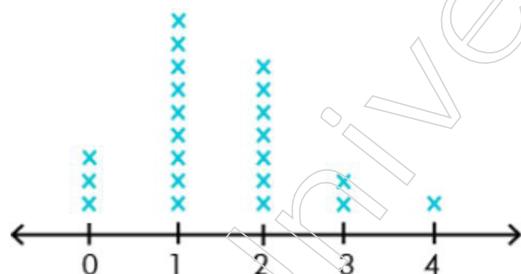
except instead of using length to convey position as a bar chart does, Cleveland dot plots make use of position instead. In his book "Elements of Graphing Data," William S. Cleveland is credited with the invention of the concept of a continuous variable. The Cleveland dot plot is helpful when working with several variables since it does not need the axis to start at zero and so enables the use of a log axis. This makes the Cleveland dot plot an attractive option.

### Wilkinson Dot Plot

The Wilkinson dot plot presents the data in a format that is quite similar to a histogram. In contrast to a histogram, which organises the data into compartments or bins, this chart displays the data as individual data points. Leland Wilkinson developed what is now known as the Wilkinson dot plot, which contributes to the standardisation of the dot plot form.

- **Line Plot**

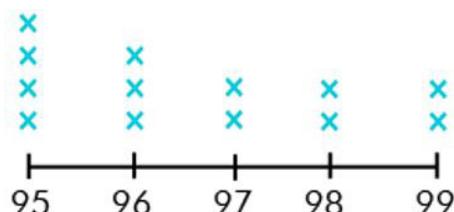
A line plot is a type of graph that can utilise Xs or any other icon to represent the number of times a response has been recorded in a given collection of data. This type of graph is called a frequency distribution plot. In most cases, the Xs should be positioned beside the replies. Line plots are also referred to as dot plots sometimes. The following is an illustration of a line plot that was derived from a survey that was carried out by students.



The number of different fruits that students consume is represented by the horizontal line. For example, in order to determine the number of students that consume two servings of fruit on a daily basis, all we need to do is count the Xs that are located above the number 2.

### How to Make a Line Plot?

In order to generate a line plot, we will first require a data set and then will need to display that data on a number line in the appropriate manner. We marked each number in the data set with an X or any other icon on top of it. If there is a certain number that shows up more than once in the data, we will indicate it with an additional X above that number. As an illustration, the line plot for the data set that contains the values 95, 95, 95, 96, 96, 96, 97, 97, 98, 98, 99 and 99 will appear as follows:



In this case, instead of using Xs to plot, we have utilised dots.

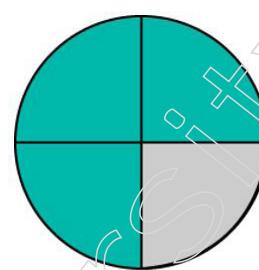
## Notes

### How to Interpret a Line Plot?

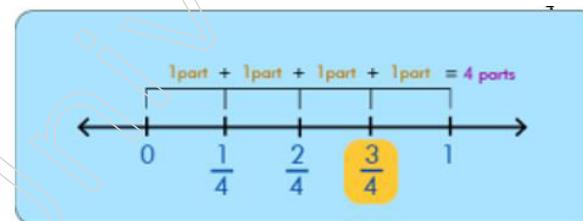
It is possible to quickly and accurately understand a line plot simply glancing at the Xs or dots that have been drawn on the line and counting them in the appropriate manner. For instance, based on the query that was just asked, we are able to determine how many times the value 96 has been found in the data collection. The value 96 has been found three times in the set of data. In a similar manner, we are able to analyse the line plot by counting the Xs and using other numbers.

### How are Fractions Depicted on a Line Plot?

The usage of fractions allows for the representation of a whole number's component components. If we are required to illustrate  $\frac{3}{4}$  using a diagram, then it will seem as follows:



The shaded parts represent  $\frac{3}{4}$ , as mentioned. Now, while representing fractions on a line plot, we will first look at fractions on a number line. We will represent  $\frac{3}{4}$  on a number line.



If you have been paying attention, you'll have seen that the number line has been cut into four equal halves. Since the fraction is a component of the whole number, we have shown it alongside the other components of the whole number, which are 1 and 0. The number line, which begins with zero and ends with one, is a representation of the many components that make up the whole number.

### Pie Charts



One style of graph that can depict the information found in a circular graph is called a pie chart. A pie chart is a sort of graphical representation of data and the slices on the pie chart indicate the relative magnitude of the data. A pie chart requires a list of

category variables and numerical variables. In this context, the phrase “pie” refers to the entire thing, while “slices” refer to individual portions of the whole.

The circular statistical graphic that is commonly referred to as a “pie chart” is also referred to as a “circle chart,” and it illustrates numerical issues by splitting the circle into sectors or portions. A proportional share of the entire is denoted by each individual sector. A pie chart is the most effective method to use at this moment for determining the component parts of something. In most situations, bar graphs, line plots, histograms and other types of graphs may be replaced with pie charts instead.

### Formula

The pie chart is an essential component of the data representation toolkit. It is composed of a variety of sections and sub-sections, with each section and sub-section of a pie chart being a distinct component of the whole (percentage). The total amount of all of the data adds up to exactly 360 degrees.

**The total value of the pie is always 100%.**

The following procedures should be followed in order to arrive at an accurate percentage for a pie chart:

- Categorise the data
- Calculate the total
- Divide the categories
- Convert into percentages
- Finally, calculate the degrees

As a result, the formula for the pie chart may be written as:

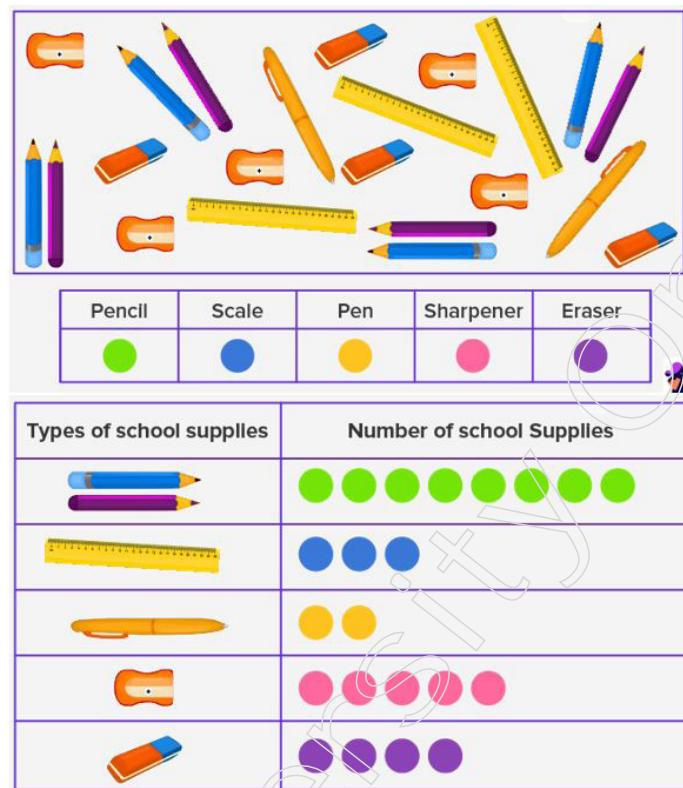
$$\text{(Given Data/Total value of Data)} \times 360^\circ$$

**Note:** Remember that it is not necessary to transform the data that has been provided into percentages unless it has been indicated to do so. We are able to do an indirect calculation to get the degrees for a set of data values and then we can construct the pie chart accordingly.

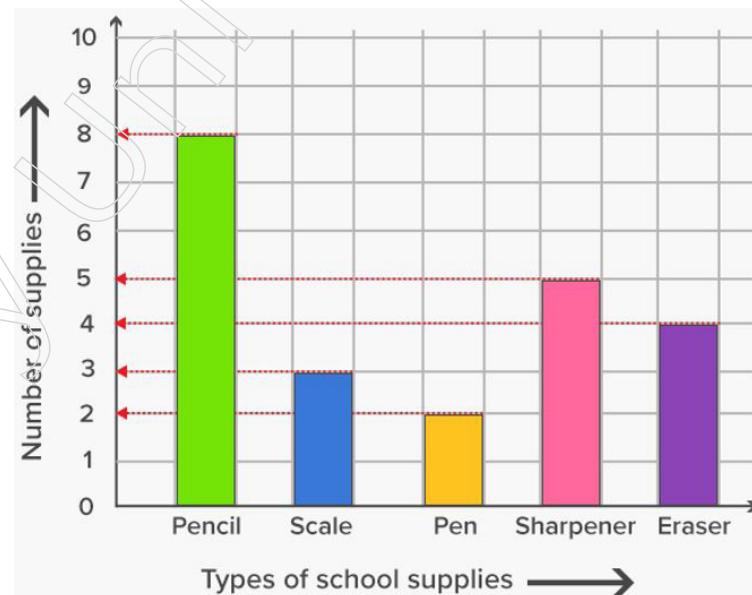
### • Graphs

A visual representation or a diagram that depicts data or values in an ordered manner may be referred to as a graph. Graphs can also be defined as diagrams. Most of the time, the points on the graph show the relationship between two or more different items. On this page, for example, we are able to depict the data given below, which is the kind and amount of school supplies utilised by pupils in a class, as a graph. To begin, we are going to count each supply and then put the data in a table using certain colours that are arranged in a specific sequence.

## Notes



A bar graph is another option for displaying the data that we have. Bars are used to illustrate how many of each of the goods are currently available. The higher the bar, the greater the amount of supplies or other objects that are being utilised.



### Types of Graphs

#### Pictograph

The act of representing information via the use of pictures is referred to as pictograph. Every image is meant to represent a given amount of objects or stuff. For illustration purposes, you may use an image of a cricket bat to show how many of such bats a certain store sold over the course of a particular week.

**Notes**

| The scale used:  = 4 cricket bats |                                                                                   |
|--------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Day                                                                                                                | Number of bats sold                                                               |
| Monday                                                                                                             |  |
| Tuesday                                                                                                            |  |
| Wednesday                                                                                                          |  |
| Thursday                                                                                                           |  |
| Friday                                                                                                             |  |
| Saturday                                                                                                           |  |
| Sunday                                                                                                             |  |

In this particular pictograph, one image of the cricket bat stands in for a total of four actual cricket bats. On Tuesday, a total of 12 bats ( $4+4+4$ ) were purchased, as shown by the graph.

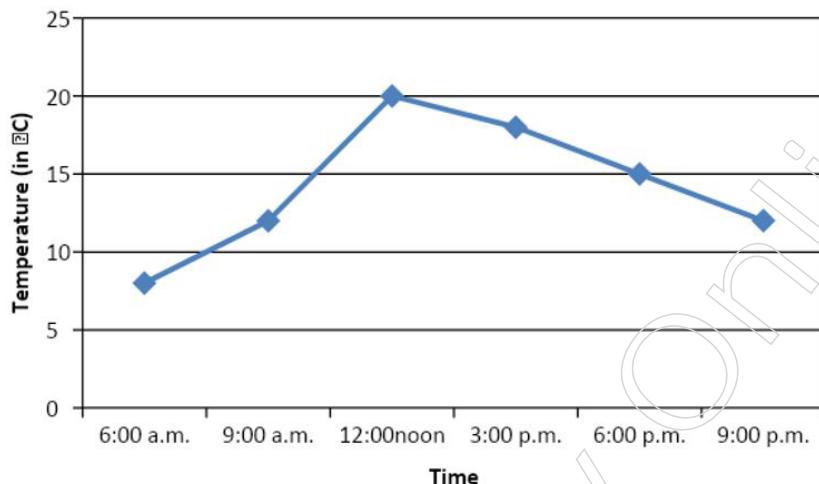
**Bar Graph**

A bar graph is a depiction of numerical data that uses rectangles (or bars) of equal width and varied height. Bar graphs are commonly used to compare different groups of data. The distance between each bar remains constant for the entirety of the chart. Both horizontal and vertical orientations are acceptable for bar graphs. There is a one-to-one correspondence between the height or length of each bar and its respective value.

**Line Graph**

A line graph depicts the changes that have occurred over a period by using dots that are connected by lines.

## Notes



### 2.2.2 Description of Data Using These Tools with Real time Example

The practise of analysing data in real time, also known as while the data is still being acquired, is referred to as real-time data analysis. It is possible to utilise it to make decisions while they are still pertinent, as opposed to waiting until after they have been made. This enables firms to direct their efforts towards ways in which they may enhance their operations, which is beneficial given that a significant proportion of internet businesses revolve on the gathering of data and the subsequent use of this information for forecasting purposes.

Real-time analytics is frequently utilised in business intelligence (BI), which is an umbrella term for software applications that examine consumer demographics or market trends in order to improve management decision-making. For instance, if you manage an online business that sells items to clients all over the world, this research would tell you exactly what your customers in each nation desire at any given moment, allowing you to adapt your inventory to meet their needs in the most effective manner possible.

The term “business intelligence” can relate not only to the process of analysing data but also to the technologies that are used for this purpose. The use of business intelligence tools can also be put to a variety of other purposes, such as the monitoring of the traffic and performance of a website, the provision of insights into the level of customer satisfaction and even the forecasting of future trends with the assistance of machine learning algorithms.

Business intelligence (BI) is a big word, but the premise behind it is straightforward: BI technologies enable you to make better decisions. If you have the right data and analysis, you can efficiently optimise your processes, which can also boost client happiness and lead to an increase in income.

#### How Does It Work?

As the data are examined as they are collected, the findings may be obtained instantly. It is essential to emphasise the fact that real-time analytics does not call for the development of new technologies. You only need a computer and connection to the Internet in order to run software like R or Python across your local network and do

real-time analysis on the data sets that are being received. An analyst can undertake analysis on their own at their workstation, or it can be done automatically by a machine learning model running on a server in your office or remotely somewhere else.

When the data has been processed and evaluated, it may be included into a model that estimates future demand for a particular good or service. Here is when ML comes into play in the story. If you have a large amount of data from the past and how it correlates with specific factors such as weather or price changes, you can use statistical methods to determine which ones are the most important and should be included in your model. This can be done when you have a large amount of data from the past and how it correlates with specific factors.

Algorithms that are used for machine learning may be used to determine which factors are the most essential and then a model can be constructed that makes use of those variables to determine how many things will be sold. The operation of Amazon's platform for cloud-based predictive analytics looks like this. On its website, it is said that "The platform ingests your data, applies ML algorithms and gives insights."

### **Why is It Important?**

Your ability to learn from your data will increase in proportion to the level of detail it contains. If you investigate the sales of a certain product over time, for instance, you could find that the demand for water bottles is highest during the summer months, when temperatures are high and people are on the move. This is because people tend to drink more water when it is hot outside.

If the only information you have is the number of bottles that were sold during each week of the previous year, it is possible that you will not be able to draw any conclusions about how the weather or pricing influences sales. You might examine past data from years with conditions that were comparable to make a prediction about how much demand would grow if none of these variables changed; but you would still be lacking some essential information.

You may better understand your consumer base and establish plans to cater to their requirements with the assistance of real-time analytics. You will also be able to gain a better understanding of the dynamics of your existing client base, which will enable you to foresee future trends and make decisions in accordance with those predictions.

The real-time analysis gives companies the ability to swiftly react to shifts in the market or in the level of competition. They are able to see chances for expansion and develop new goods or services on the basis of these insights into the desires and requirements of their customers—before their rivals follow suit.

### **Examples**

Real-time data processing may be utilised in a number of settings and applications. In the following, we will examine several instances from the following three primary business areas: fraud detection, marketing and providing support to customers.

#### **a) Fraud detection**

The usage of fraudulent user activity within your company or organisation can be uncovered through the use of real-time analytics. This can include tracking the

## Notes

purchases that customers make or tracing the customers' IP addresses back to the geographic location where they may be committing fraud against the company. For example, this could include an attempt to steal credit card information or to engage in an identity theft scam. Real-time analysis may also be employed by a corporation in order to monitor the high-risk activities of its clients. This can assist it identify what measures should be done next, such as a credit card business decreasing a client's credit limit or even cancelling the customer's account if it is determined that the consumer is not paying their bills.

### b) Marketing

Retailers may improve their marketing efforts with the use of real-time analytics by first determining which goods customers look at the least frequently and which they look at the most frequently and then modifying their sales strategy in accordance with those findings. For example, if a large number of individuals are looking at shirts but not purchasing them in significant quantities (indicating that those shirts might need better placement). Then, retailers would be able to reorganise their stock so that shirts would be more prominently displayed throughout the store rather than being buried under other items of clothing. This would make it less likely for customers to become disoriented while they were browsing through the same sections multiple times.

This is just one illustration of how merchants may make use of data to learn more about the buying patterns and preferences of their consumers, which in turn enables them to better satisfy those requirements. Mobile analytics software enables businesses to handle data streams in a more effective manner, allowing for informed decisions to be made about inventory management and product placement. This, in turn, means that merchants will have a better grasp of what it is that customers desire.

### c) Customer service

Monitoring client chats in real time regarding your products or services is another use for real-time analytics. You may then put this knowledge to use by conducting advertising and marketing campaigns on websites that are well-liked by clients. These advertisements reach a greater number of prospective clients, some of whom may be unaware of the presence of your web business. This is something that may be very helpful for businesses that offer items or services online, such as an e-commerce website. Companies may figure out what their consumers want by using the information they gather from social media, which in turn helps them develop better products and enhance their pricing tactics. Corporations can utilise this information to collect it.

## Topic 2.3 Basic Data Science Process

### Introduction

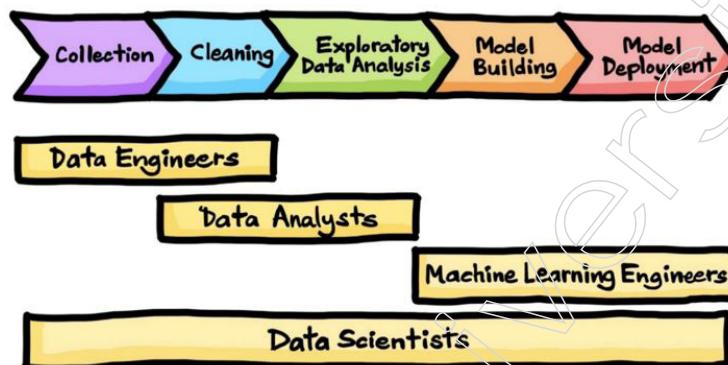
Data science is a subfield of computer science that focuses on extracting useful insights from large amounts of data using a combination of domain experience, programming abilities and understanding of mathematics and statistics. Data scientists develop artificial intelligence (AI) systems by applying machine learning algorithms to a variety of data types, including numbers, text, photos, video and audio, among other things. These AI systems are able to execute jobs that would normally need human intellect. On the other hand, these technologies produce insights, which analysts and business users may then convert into measurable value for the company.

The combination of different tools, methodologies and technologies into a single field is what makes data science such an essential field today. The development of gadgets that are capable of automatically collecting and storing information has resulted in a deluge of data for today's businesses, which are struggling to keep up. In the disciplines of e-commerce, medical and finance, as well as every other element of human existence, online platforms and payment gateways acquire more data than ever before. We have text, audio, video and image data available in large amounts.

### 2.3.1 Overview of Data Science Process: Defining its Goal

The process of data science refers to a methodical way of approaching the resolution of a data issue. It gives a systematic framework for expressing your problem as a question, choosing how to answer it and then delivering the solution to stakeholders after you have decided how to solve it.

#### Data Science Life Cycle



The data science life cycle is another name for the process that data scientists go through. Both phrases describe a workflow process that begins with gathering data and finishes with installing a model that will ideally answer your questions. The expressions may be used interchangeably with one another and both phrases explain the workflow process. The following are the steps:

#### (a) Framing the Problem

The first phase in the process of performing data science is known as "understanding and defining the problem." You will be able to develop an efficient model with the aid of this framing, which will result in beneficial effects for your business.

#### (b) Collecting Data

The next thing that has to be done is to acquire the appropriate assortment of data. It is impossible to produce significant results without collecting data of a high quality and specificity, as well as the procedures to do so. It is quite likely that you will be required to extract the data and convert it into a format that is useable, such as a CSV or JSON file, due to the fact that the majority of the approximately 2.5 quintillion bytes of data that are produced every day originate in unstructured formats.

#### (c) Cleaning Data

The vast majority of the unstructured, irrelevant and unfiltered data that you acquire

## Notes

during the phase known as collection will not be filtered. As inaccurate data leads to inaccurate findings, the accuracy and effectiveness of your analysis will strongly depend on the quality of the data you provide.

Cleaning data eliminates duplicate and null values, corrupt data, inconsistent data types, incorrect entries, missing data and poor formatting. Finding and fixing errors in your data is one of the most time-consuming parts of the modelling process, but doing so is necessary in order to construct reliable models.

### (d) Exploratory Data Analysis (EDA)

You are now in a position to start an exploratory data analysis since you have a substantial quantity of data that is well-organised and of good quality (EDA). Discovering meaningful insights that may be used to the subsequent stage of the data science lifecycle is made possible by employing an efficient EDA system.

### (e) Model Building and Deployment

Following this, you will be responsible for the actual data modelling. Machine learning, statistical modelling and algorithmic analysis will all be utilised at this stage of the process in order to glean insights and forecasts of high value.

### (f) Communicating Your Results

In the last step, you will report your results to the various stakeholders. To do this, every data scientist must expand their skill set in the area of visualisation. Your stakeholders are more concerned with what your findings represent for their company than they are with the intricate back-end work that was done to construct your model. In most cases, they will not be interested in this information. You should communicate your results in a way that is both clear and engaging, calling attention to the significance of those findings for strategic company planning and operation.

## Significance of Data Science Process

Every company or organisation that implements a data science methodology will reap several benefits. Also, it is now of the utmost significance for the achievement of success in any firm. Incorporating a procedure from the field of data science into your standard practise of data collecting should be strongly considered for the following reasons:

1. Produces better outcomes and results in increased production.

There is no question that a corporation or business that possesses data or has access to data enjoys a competitive edge over those that do not. It is possible to process data in a variety of formats to gather the information that the organisation requires and to assist it in making sound decisions. Decisions may be made and company executives can feel confident in those judgements, thanks to the support that statistics and details provide, when a data science methodology is used. This provides the organisation with an advantage over its competitors and boosts overall productivity.

2. The process of creating reports is made easier

The collection of values and the creation of reports based on those values is nearly always accomplished through the usage of data. After the data has been suitably

processed and placed into the framework, it can be simply accessible without any fuss with the press of a button, which makes the preparation of reports a matter of only minutes rather than hours.

### 3. Speedy, Accurate and More Reliable

It is of the utmost significance to make certain that the amassing of data, facts and statistics is carried out in a timely manner and without committing any errors. If a data science technique is used to the data, there is a very little to almost non-existent probability of errors or mistakes occurring. This ensures that the procedure that comes after it may be carried out with a higher degree of precision. In addition, the procedure yields superior outcomes. It is not at all unusual for numerous different rivals to possess the same data. In this scenario, the business that possesses the data that is both the most accurate and the most dependable has an edge.

### 4. Convenient Facilities for Storage and Distribution

When terabytes upon terabytes of data need to be saved, the location in which this must take place must be monstrous. This increases the likelihood that important information or data may be lost or misunderstood. A data science procedure will provide you with more space to store documents and complicated files, as well as the ability to label the entire dataset using a computerised system. This results in a reduction of misunderstanding and facilitates quick access to and utilisation of the data. One of the benefits of doing data science is that it results in the data being saved in digital format.

### 5. Reduced costs

The requirement to repeatedly collect and examine data may be avoided by utilising a data science method to collect and store data instead. Also, it makes it easy to create duplicates of the material that has been stored in digital format. It is now much simpler to send or transfer data for the purpose of doing research. Because of this, the overall cost to the organisation is decreased. It protects the data that would otherwise be at risk of being lost in papers, which in turn supports a decrease in costs. Using a data science approach can help decrease losses that are incurred as a result of a lack of specific data. The ability to make informed and confident judgements, which in turn leads to cost savings, is made possible by data.

### 6. Risk-free and risk-free

Having information digitally saved after going through a data science process makes the information far more secure. The fact that the value of data tends to rise over time is one factor that has contributed to the increased frequency with which it is stolen. When the data has been processed, it is next encrypted and protected from illegal access by a variety of software programmes. These programmes work in tandem to ensure that your data is safe.

## 2.3.2 Retrieving the Data, Data Preparation-Exploration, Cleaning and Transforming Data

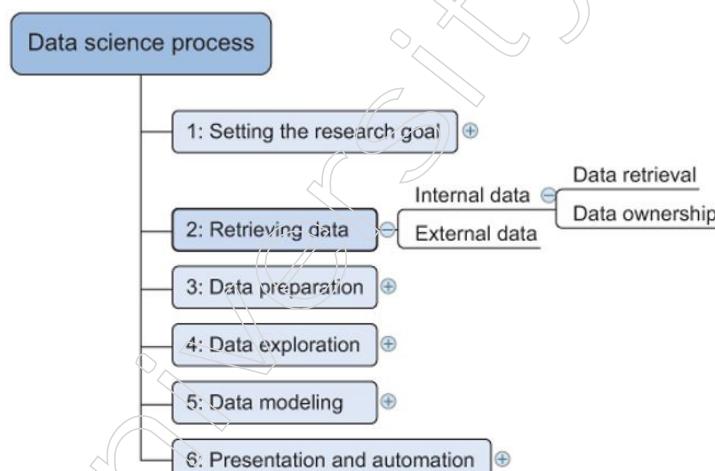
### • Data Retrieval

In most cases, retrieving data requires the creation and execution of instructions or queries specifically designed for the purpose of data retrieval or extraction from a

## Notes

database. The database will search for and obtain the desired information based on the query that is submitted to it. Applications and software make use of a variety of queries to get data in a variety of forms. Data retrieval can involve the retrieval of vast volumes of data, which are often presented in the form of reports. This is in addition to the retrieval of basic or smaller data.

The collection of the necessary data is one of the phases involved in data science (Figure 1). Most of the time, you will not be required to participate in this stage of the process; nevertheless, there may be instances in which you will be required to travel to the location of the data collecting and develop the procedure yourself. A great number of businesses will have already gathered and stored the information for you and the information that they do not have may frequently be purchased from other parties. Do not be hesitant to explore for data outside of your company since an increasing number of organisations are making data of any type, including high-quality data, publicly available for use by the public and by businesses.



**Figure 1: Retrieving data**

There are a variety of formats in which data can be saved, from plain text files to tables organised in a database. Obtaining all the necessary data should now be the focus of your efforts. This may be tough to do and even if you are successful, the data is frequently like an unpolished diamond; for it to be of any value to you, it needs to be polished.

### Start with Data Stored Within the Company

Your first order of business should be to evaluate the usefulness and precision of the information that is at your disposal inside your organisation. As most businesses already have a system in place for the management of essential data, a significant portion of the cleaning job may already have been completed. This information may be kept in formal data repositories like as databases, data marts, data warehouses and data lakes, which are all tended to by groups of knowledgeable IT specialists. The major objective of a database is the storing of data, but the data reading and analysis capabilities of a data warehouse are the driving forces behind its development. A data mart is a section of a data warehouse that is dedicated to the needs of a certain department or division of a company. In contrast to data warehouses and data marts, which store data after it has been pre-processed, data lakes store data in its original,

unprocessed version. But there is a potential that your data is still stored in Excel files on the computer of a specialist in the relevant field.

Even inside your own firm, it might be difficult to track down the data you need at times. When businesses expand, the data they collect gets dispersed across a variety of locations. If employees move across departments or leave the organisation entirely, the information they formerly knew might become fragmented. Documentation and metadata are not often a delivery manager's top concern, therefore it is feasible that you'll need to develop some talents comparable to those of Sherlock Holmes in order to uncover all of the missing pieces.

Another challenging undertaking is gaining access to the data. Because organisations are aware of the importance of data as well as its sensitivity, policies are frequently put in place to ensure that users only have access to the data that is necessary to them. These rules have the effect of erecting barriers, known as Chinese walls, both physically and virtually. In most countries, having these "walls" around client data is not only required but also strictly enforced. Imagine if every employee at a credit card firm had access to information on your spending patterns; this is one of the reasons why this is the case. The process of gaining access to the data might take some time and include the politics of the firm.

### **Do not be Afraid to Shop Around**

If the data you need is not readily available within your business, you should explore beyond its boundaries. There are several firms whose primary focus is the gathering of useful information. For example, Nielsen and GFK are quite well recognised in the retail business for their work in this area. You may improve the services and ecosystem that other firms provide by using data that is provided by other companies. This is the situation with social media platforms like Twitter, LinkedIn and Facebook. Even while some businesses believe data to be an asset with a value greater than that of oil, an increasing number of governments and organisations are making their data freely available to the public online. The organisation that generates and oversees the management of this data can have a significant impact on the data's overall quality. They discuss a wide variety of issues, such as the number of accidents that take place or the quantity of drug usage that occurs in a certain location in addition to the demographics of that area. This data is useful not just when you want to augment private data but also when you are practising your data science abilities at home because of its convenience. The table 1 only includes a small sample of the ever-increasing number of open-data suppliers.

**Table 1. A list of open-data providers that should get you started**

| <b>Open data site</b>                                                   | <b>Description</b>                                                                                         |
|-------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| Data.gov                                                                | The home of the US Government's open data                                                                  |
| <a href="https://open-data.europa.eu/">https://open-data.europa.eu/</a> | The home of the European Commission's open data                                                            |
| Freebase.org                                                            | An open database that retrieves its information from sites like Wikipedia, MusicBrains and the SEC archive |
| Data.worldbank.org                                                      | Open data initiative from the World Bank                                                                   |
| Aiddata.org                                                             | Open data for international development                                                                    |
| Open.fda.gov                                                            | Open data from the US Food and Drug Administration                                                         |

## Notes

### Do data quality checks now to prevent problems later

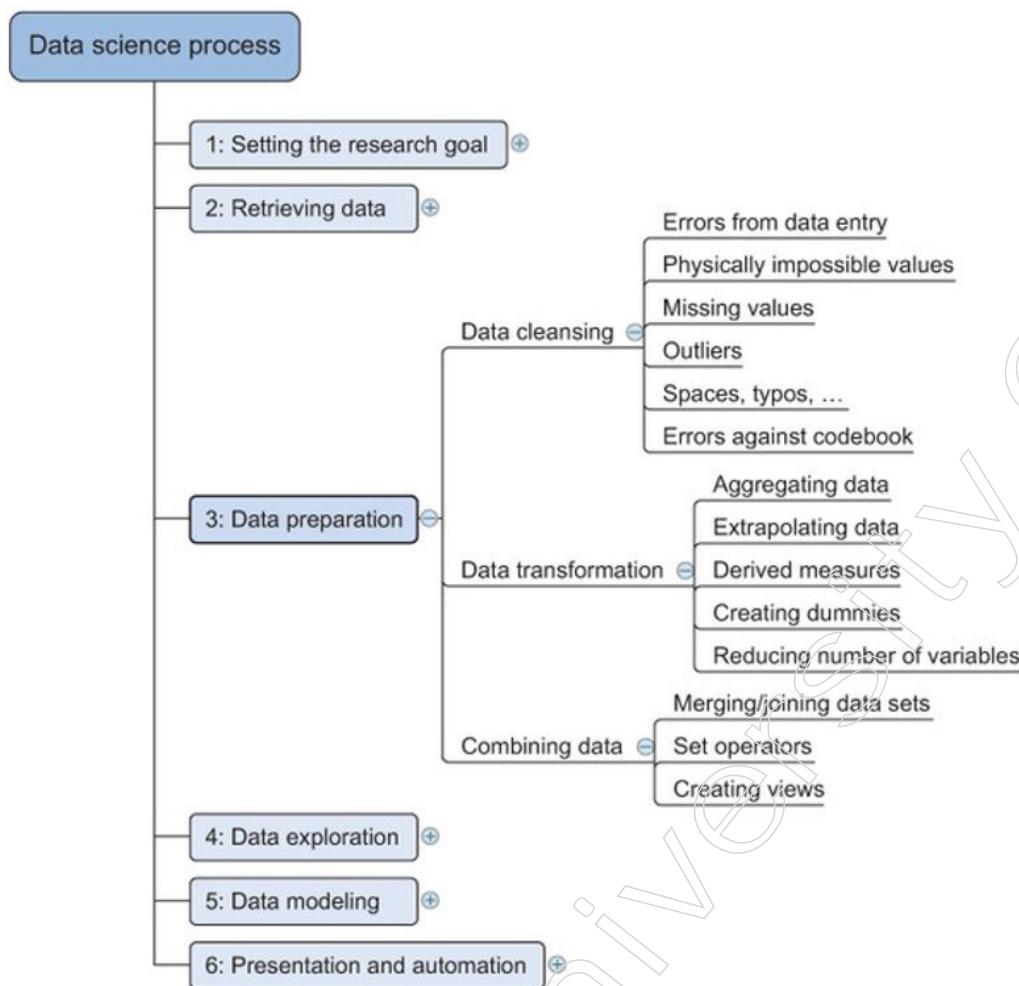
Check the quality of the data now to avoid difficulties in the future. You should plan on devoting a considerable percentage of your project's time to correcting and cleaning the data, perhaps as much as 80 percent. In the process of doing data science, the initial inspection of the data will take place when you retrieve the data. It should not be too difficult to identify most of the errors that crop up during the phase in which you are collecting the data; but, if you are too reckless, you will wind up having to spend a lot of time fixing data problems that might have been avoided during the data import phase.

Throughout the steps of importing and preparing the data, as well as exploratory analysis, you will explore the data. The distinction lies in the objectives of the research as well as its breadth. Throughout the process of retrieving data, you must verify that the data matches the data found in the source document and examine your collection of data to ensure that you have the appropriate data types. This should not take too much time; you will know you are done when you have sufficient proof showing that the data is comparable to the data found in the source text. Throughout the process of data preparation, a more thorough inspection is performed. If you did a decent job with the step before this one, the problems that you detect now are also present in the document that was used as a source. The substance of the variables is where the focus should be, you want to eliminate typos and other problems that may have occurred during data entry and bring the data up to a common standard across all the data sets. You may, for instance, change USQ to USA and United Kingdom to UK. As you progress through the exploratory phase, you will move your attention to what you can learn from the data. Now that you have assumed that the data are free of errors, you may investigate the statistical aspects, such as distributions, correlations and outliers. You will find yourself returning to these periods rather frequently. For instance, if you find outliers while in the exploratory phase, they can indicate that there was a mistake with the data input. Now that you understand how the data's quality is enhanced throughout the process, we will proceed to examine the stage of data preparation in further detail.

#### • **Data Preparation-Exploration, Cleaning and Transforming Data**

There is a good chance that "a diamond in the rough" will be found in the data that was obtained through the data retrieval phase. Your job right now is to clean it up and get it ready to utilise for the modelling and reporting portion of the process. If you do so, the performance of your models will improve and you will waste less time attempting to resolve unusual output. This makes it an extremely vital step to take. The adage "garbage in, rubbish out" cannot be repeated nearly enough: both sides contribute to the problem. As your model requires the data to be in a particular format, the process of data translation will continuously be an issue. It's a good practise to go back as early as possible in the process and fix any data problems you find. Yet, this isn't always achievable in a practical scenario, so you'll need to make adjustments to your software to account for that.

The most frequent steps to perform throughout the period of data purification, integration and transformation are depicted in Figure 2.

**Figure 2: Data preparation**

- **Cleansing data**

The subprocess of data science known as “data cleaning” focuses on eliminating inaccuracies from one’s data to make that data a more accurate and consistent reflection of the processes that it was derived from. This allows the data to be used more effectively.

When we say, “true and consistent representation,” we are giving the impression that there are at least two different kinds of faults. The first kind of error is an interpretation error, which can occur when, for example, you assume about a value included in your data, such as stating that a person’s age is higher than 300 years. The second kind of mistake indicates that there are conflicts either between the data sources or against the standardised values used by your firm. Putting “Female” in one table and “F” in another when they both convey the same thing: that the individual is female, is an illustration of the type of blunder that falls into this category. Another illustration of this would be the fact that one table uses pounds, while the other table uses dollars. This list cannot be thorough since there are too many possible faults; nonetheless, table 2 provides an overview of the sorts of problems that may be found with straightforward tests; these errors are sometimes referred to as the “low hanging fruit.”

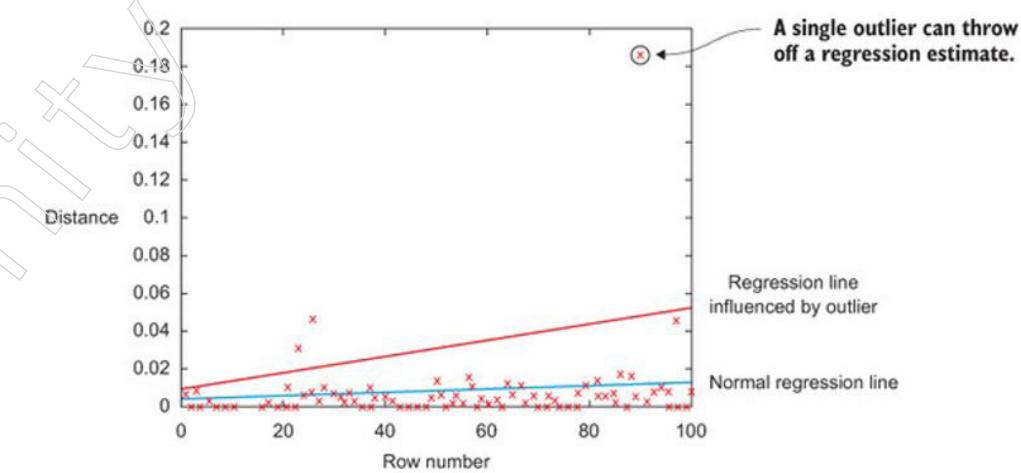
**Notes**

## Notes

**Table 2. An overview of common errors**

| <b>General solution</b>                                                                   |                                                                       |
|-------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|
| Try to fix the problem early in the data acquisition chain or else fix it in the program. |                                                                       |
| <b>Error description</b>                                                                  | <b>Possible solution</b>                                              |
| Errors pointing to false values within one data set                                       |                                                                       |
| Mistakes during data entry                                                                | Manual overrules                                                      |
| Redundant white space                                                                     | Use string functions                                                  |
| Impossible values                                                                         | Manual overrules                                                      |
| Missing values                                                                            | Remove observation or value                                           |
| Outliers                                                                                  | Validate and, if erroneous, treat as missing value (remove or insert) |
| Errors pointing to inconsistencies between data sets                                      |                                                                       |
| Deviations from a code book                                                               | Match on keys or else use manual overrules                            |
| Different units of measurement                                                            | Recalculate                                                           |
| Different levels of aggregation                                                           | Bring to same level of measurement by aggregation or extrapolation    |

Finding and identifying data mistakes may require you to employ more complex approaches at times, such as basic modelling; diagnostic charts may be extremely enlightening. For instance, in figure 3, we utilise a measure to find data points that do not seem to fit in with the rest of the picture. A regression is performed so that we may familiarise ourselves with the data and determine the impact that various observations have on the regression line. It is possible that there is a mistake in the data if a single observation has an excessive amount of effect; yet it is also possible that this is a legitimate point. At the stage of data purification, however, these more complex procedures are rarely utilised and certain data scientists frequently see them as an unnecessary excess.



**Figure 3. The encircled point has a significant impact on the model and should be investigated because it may direct you to a region for which you lack sufficient data, it may indicate that there is an error in the data, or it may simply be a valid data point. However, it is possible for it to do either of these things.**

After providing an overview of the situation, it is time to provide a more in-depth explanation of these faults.

## Notes

### Data entry errors

The methods of data gathering and data input both have a high potential for mistake. They frequently need the participation of a person and as people are fallible beings, it is not uncommon for them to make a typo or to become distracted for a split second, therefore introducing a flaw into the process. Yet, the accuracy of the data that is gathered by machines or computers cannot be guaranteed. Some errors are caused by human carelessness, while others are the result of a malfunctioning system or piece of technology. Transmission failures and defects that occur during the extract, convert and load phases are all examples of faults that may be attributed to machines (ETL).

When working with relatively modest data sets, you can verify each value by hand. Tabulating the data using counts is an effective method for finding mistakes in the data even when the variables being studied do not contain a large number of classes. When you have a variable that can only take two values, such as “Good” and “Bad,” you may make a frequency table to determine whether or not those are the only two values that are really being used in the system. Within the context of table 3, the values “Godo” and “Bade” indicate that there was an error in at least 16 of the occurrences.

**Table 3. Detecting outliers on simple variables with a frequency table**

| Value | Count   |
|-------|---------|
| Good  | 1598647 |
| Bad   | 1354468 |
| Godo  | 15      |
| Bade  | 1       |

Most errors of this type are easy to fix with simple assignment statements and if-then-else rules:

```
if x == "Godo":
 x = "Good"

if x == "Bade":
 x = "Bad"
```

### Redundant whitespace

Yet, much like other repetitive characters, whitespaces can create mistakes even if they are notoriously difficult to see. Who among us hasn't seen a problem in a project that was caused by whitespaces at the end of a string that cost them a few days' worth of work? After requesting that the software connect the two keys, you check the output file and discover that certain observations are missing. You spend many days searching through the code before eventually coming upon the error. The next step is the most difficult one, which is to explain the delay to the project's stakeholders. The cleaning that was supposed to take place during the ETL process was poorly carried out and as a result, the keys in one table had a whitespace at the end of a string. This resulted in a mismatch of keys, such as “FR” – “FR,” which led to the elimination of observations that could not be matched.

## Notes

Fixing superfluous whitespaces is fortunately not too difficult in most programming languages, provided that you are aware of where to look for them. They all have string routines that will get rid of the leading and trailing whitespaces in a string. For instance, if you want to get rid of leading and trailing spaces in Python code, you may use the method called `strip()`.

### Fixing Capital Letter Mismatches

Capital letter mismatches are frequent. The terms “Brazil” and “brazil” are treated differently by the majority of computer languages. In this scenario, the issue can be resolved by making use of a method that converts both strings to lowercase, such as `.lower()` in Python. It is expected that `“Brazil”.lower() == “brazil”.lower()` will return true.

### Impossible values and sanity checks

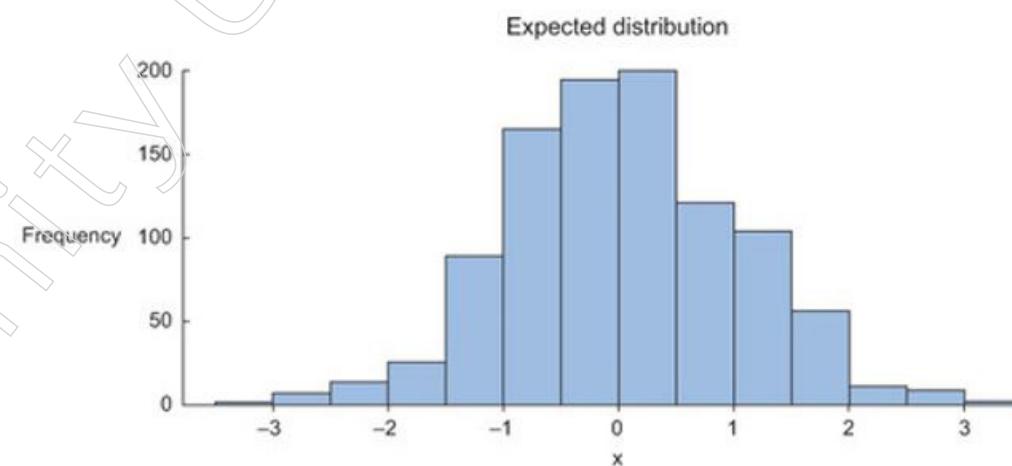
In addition to integrity checks, sanity checks are another important sort of data check. At this step, you compare the value to those that are literally or theoretically impossible, such as someone with a height more than 3 metres or an age greater than 299 years. Checks for sanity can be explicitly articulated with rules like follows:

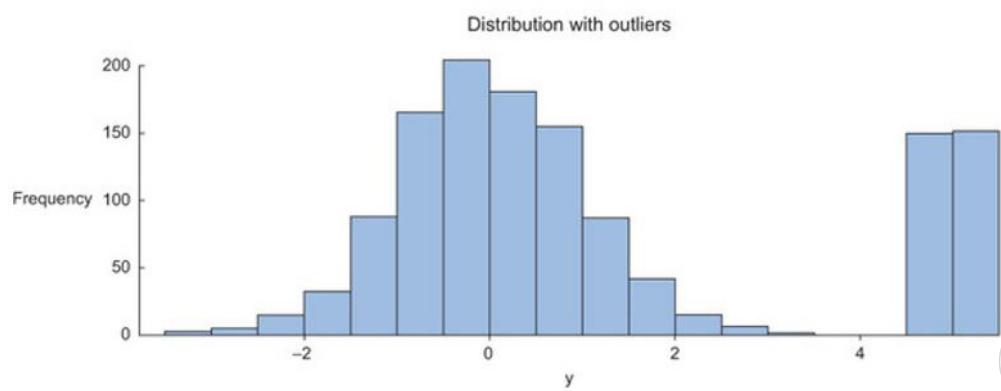
Top of Form

`check = 0 <= age <= 120`

### Outliers

An observation is considered to be an outlier if it deviates significantly from the other observations or, more precisely, if it adheres to a distinct line of reasoning or method of generation in comparison to the other observations. Using a graphic or table that lists the minimum and maximum values is the method that provides the most straightforward results when searching for outliers. Figure 4 illustrates an example of this type of thing.





**Figure 4.** The use of distribution plots can assist in the identification of outliers as well as the comprehension of the variable.

The figure on top displays no outliers, whereas the plot on the bottom indicates probable outliers on the upper side when a normal distribution is anticipated. Normal distribution, often known as Gaussian distribution, is the most prevalent distribution in the natural sciences. It demonstrates that the majority of instances occur close to the distribution's mean and that their frequency decreases as they go away from it. Assuming a normal distribution, the high numbers on the bottom graph may represent outliers. As we saw before with the example of regression, outliers can have a significant impact on your data modelling, therefore study them first.

### Dealing with missing values

Missing values are not inherently incorrect, but they must be handled independently; certain modelling approaches cannot accommodate missing values. These might be an indication that something went wrong with your data gathering or that an ETL process issue occurred. Typical strategies data scientists utilise are described in table 4.

**Table 4.** An overview of techniques to handle missing data

| Technique                                                    | Advantage                                                                              | Disadvantage                                                                                                                                   |
|--------------------------------------------------------------|----------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Omit the values                                              | Easy to perform                                                                        | You lose the information from an observation                                                                                                   |
| Set value to null                                            | Easy to perform                                                                        | Not every modeling technique and/or implementation can handle null values                                                                      |
| Impute a static value such as 0 or the mean                  | Easy to perform You don't lose information from the other variables in the observation | Can lead to false estimations from a model                                                                                                     |
| Impute a value from an estimated or theoretical distribution | Does not disturb the model as much                                                     | Harder to execute You make data assumptions                                                                                                    |
| Modeling the value (nondependent)                            | Does not disturb the model too much                                                    | Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions |

## Notes

### Dealing with Missing Values

Missing values are not inherently incorrect, but they must be handled independently; certain modelling approaches cannot accommodate missing values. These might be an indication that something went wrong with your data gathering or that an ETL process issue occurred. Typical strategies data scientists utilise are described in table 4.

Which approach to employ and when depends on your specific situation. If, for example, you do not have any extra observations, skipping one is probably not an option. If the variable can be characterised by a stable distribution, it is possible to impute using this information. But, perhaps a missing value signifies “zero”? This can occur in sales, for example: if no promotion is applied to a customer’s cart, that customer’s promo is absent, but it’s also probable 0 and no price reduction.

### Deviations from a Code Book

Set operations can be used to detect faults in bigger data sets against a code book or against specified values. A code book is a sort of metadata that describes your data. It includes the number of variables per observation, the number of observations and the meaning of each variable encoding. (For example, “0” equals “negative”, whereas “5” = “extremely positive”). A code book also specifies whether the data being seen is hierarchical, graph-based, or other.

You examine the values present in set A but absent in set B. These are values that require modification. It is no accident that sets will be the data structure we employ when writing programming. It is a good practise to give additional attention to your data structures; doing so can save time and enhance the performance of your software.

If you need to compare numerous values, it is best to put them from the code book into a table and use a difference operator to see whether there is a disparity between the two tables. So, you may immediately benefit from the power of a database.

### Different Units of Measurement

When integrating two data sets, the respective units of measurement must be considered. This is the case, for instance, when examining global petrol prices. To achieve this, you collect data from several data suppliers. Some data sets may include prices per gallon, while others may include prices per litre. A straightforward conversion will suffice in this instance.

### Various Degrees of Aggregation

Having several aggregate levels is akin to having multiple measurement kinds. This is illustrated with a data collection having data per week as opposed to one holding data per work week. This sort of inaccuracy is typically straightforward to discover and it may be corrected by summarising (or enlarging) the data sets.

After correcting the data inaccuracies, you mix data from several sources. But, before we get into this issue, we will take a brief diversion to emphasise the necessity of data cleansing as early as feasible.

### Correct Errors as Early as Possible

A best practise is to correct data mistakes as early as feasible in the data gathering chain and to repair as little as possible within the programme while addressing the source of the issue. Organisations invest millions of dollars in the retrieval of data in an effort to make more informed judgements. Error-prone and involving several phases and teams in large organisations, the data collecting process is susceptible to mistakes.

As data is gathered, it must be sanitised for several reasons:

- Not everyone recognises data irregularities. Decision-makers may make expensive errors when relying on information derived from programmes that fail to compensate for inaccurate data.
- If mistakes are not rectified early in the process, data cleaning will need to be performed on every project that utilises the data.
- Data mistakes may indicate that a business process is not operating as intended. For instance, both writers previously worked for a shop, where they devised a couponing system to attract more customers and increase profits. During a project involving data science, we uncovered clients who misused the couponing system and made money while shopping groceries. The purpose of the couponing system was to encourage cross-selling, not to provide free things. Nobody in the corporation was aware that this error had cost the business money. In this instance, the data was not technically incorrect, but the outcomes were unexpected.
- Data errors may indicate faulty hardware, such as damaged transmission lines and faulty sensors.
- Errors in data might indicate problems in software or in the integration of software that could be detrimental to the business. Over the course of a minor project at a bank, we learned that two software programmes utilised distinct local configurations. This caused issues with numbers bigger than one thousand. One app interpreted the number 1.000 as one, while another app interpreted it as one thousand.

In a perfect world, data would be rectified as soon as it is recorded. However, a data scientist does not always have a say in data collecting and merely instructing the IT staff to solve certain issues may not be sufficient. If you are unable to fix the data at its source, you will have to manage it within your code. Correction of errors is not the end of data manipulation; you must also merge your incoming data.

Always maintain a backup copy of your original data (if possible). Occasionally, when you begin cleaning data, you will make mistakes, such as incorrectly imputing variables, deleting outliers with fascinating extra information, or modifying data because of an initial misreading. If you keep a duplicate, you get a second chance. This is not always practicable for “flowing data” that is modified upon arrival and you must allow a period of adjustment before you can use the captured data. One of the most challenging tasks is not the data purification of individual data sets, but rather merging disparate sources into a coherent whole.

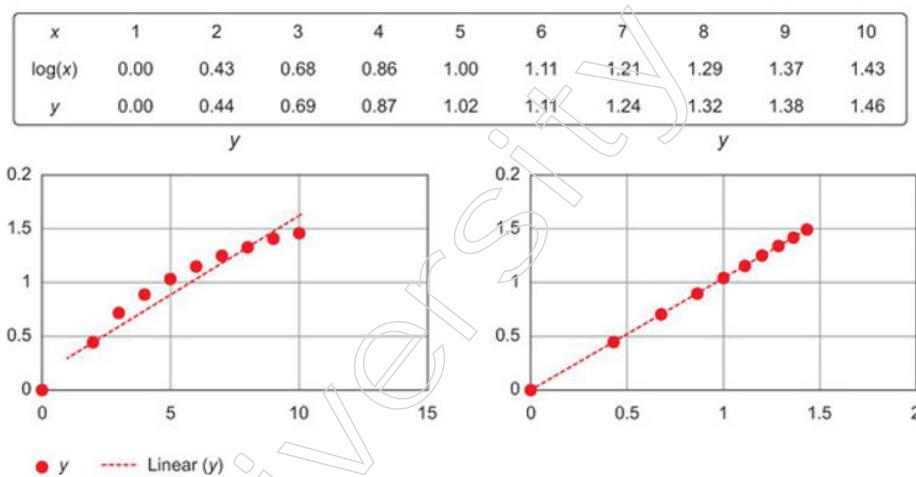
## Notes

- Transforming Data

Some models demand that their data be in a certain format. When you have cleaned and integrated the data, the following step is to change the data into a format that is acceptable for data modelling.

### Converting Data

Not always are relationships between an input variable and an output variable linear. Consider a connection of the kind  $y = ae^{bx}$  as an example. Considering the logarithm of the independent variables substantially simplifies the estimate issue. Figure 5 illustrates how changing the input variables reduces the estimate issue considerably. On sometimes, you may wish to merge two variables into a single variable.



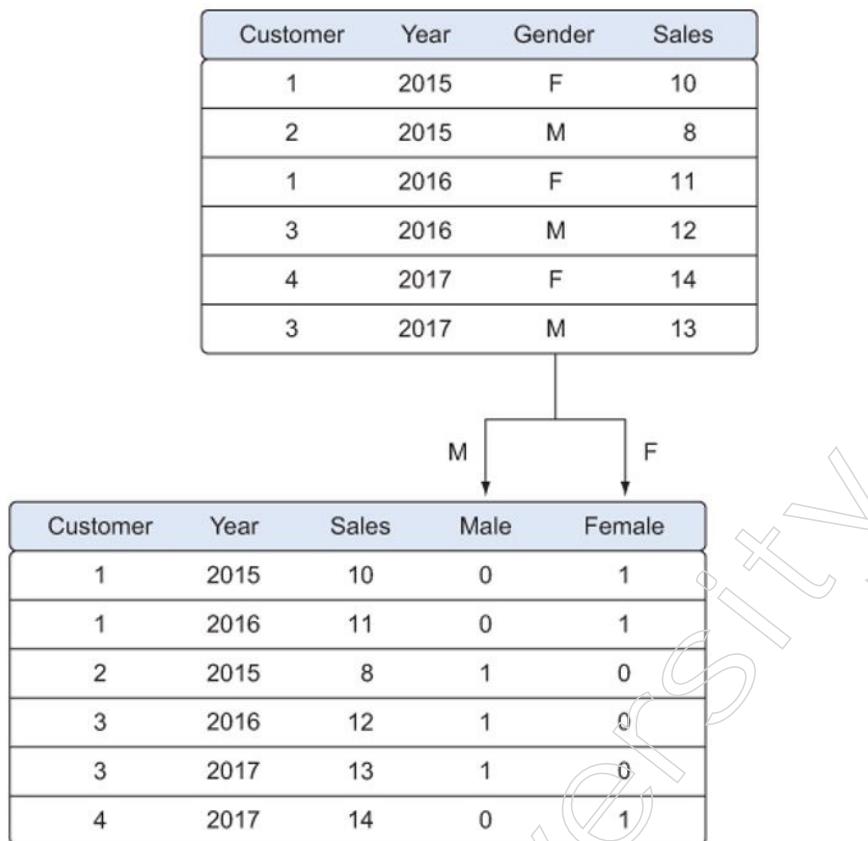
**Figure 5.** Transforming  $x$  to  $\log x$  makes the relationship between  $x$  and  $y$  linear (right), compared with the non-log  $x$  (left).

### Reducing the Number of Variables

Sometimes it is necessary to minimise the number of variables in a model since they do not provide new information. A model with too many variables is difficult to manipulate and certain approaches perform poorly when they are overloaded with too many input variables. All strategies based on Euclidean distance, for instance, only work well up to 10 variables.

### Variables Transformed into Dummies

Variables are capable of being transformed into dummy variables (figure 6). false (1) or true (1) are the only valid values for dummy variables (0). They imply the absence of a categorical effect that may explain the finding. In this situation, you will create distinct columns for the classes contained in a single variable and indicate their presence with a 1 and absence with a 0. An example would be transforming the Weekdays column into the columns Monday through Sunday. You use an indicator to indicate if the observation occurred on a Monday; you place a 1 on Monday and a 0 otherwise. Converting variables into dummies is a modelling approach that is popular among economists, but not limited to them.



**Figure 6. Turning variables into dummies is a data transformation that breaks a variable that has multiple classes into multiple variables, each having only two possible values: 0 or 1.**

### 2.3.3 Building the Model

#### Model Building

At this phase, the data science team must construct data sets for training, testing and production. These data sets enable data scientists to create and train an analytical approach, while reserving some data for testing the model. The team creates datasets for use in testing, training and production. In addition, the team constructs and executes models based on the work completed in the model planning phase. The team also analyses if its current tools are enough for running the models, or whether a more robust environment is required for executing models and processes (Example – fast hardware and parallel processing).

#### Step 1: Understanding Business Problem

Although this is not a phase in constructing a data science model, experts think that if data scientists do not understand the business challenge, they have no basis for building a data science model. One should be aware of the issue that data scientists are attempting to tackle. Comprehend the data science process model and the end goal of developing a data science business model. In addition, having defined, quantifiable objectives will enable data scientists to quantify the ROI of the data science project, as opposed to merely deploying a proof of concept that will be kept aside later.

## Notes

### Step 2: Data Collection

Once data scientists are aware of the problem they are attempting to answer, they gather data. Data collection involves acquiring both organised and unstructured data that is meaningful. Notable data repositories include Dataset Search Engines, Kaggle, NCBI, UCI ML Repository and others. Unless they obtain data that is relevant to the business challenge, data scientists spend the majority of their time sifting data.

### Step 3: Prepare Data

Prepare Data Once relevant data has been gathered, data scientists must prepare the data in order to train the data science model. Data preparation includes data cleansing, data aggregation, data labelling, data transformation, etc. Methods for data preparation include:

- Standardize formats across diverse data sources
- Reduce deduplication data
- Remove erroneous data
- Enhance and supplement data
- Normalize or standardise data to bring it into structured ranges
- Split data into testing and validation sets.

Consider that data cleansing and preparation is a time-consuming process. Yet, it is also an essential stage in creating data science models. The effort spent cleansing data has undeniably significant returns.

### Step 4: Analyse Patterns in Data

Following data cleansing, data scientists have important and relevant data for creating models in data science. The following phase is to recognise patterns and trends in data. At this level, Micro strategy and Tableau are useful tools. Data scientists must create an understandable dashboard and identify major data patterns. Data scientists would be aware of the underlying causes of business issues. In the case of pricing features, for instance, they would be aware of all pertinent information, such as if the price fluctuates, why, when, etc.

### Step 5: Training Model Features

Now that data scientists have access to high-quality data and knowledge on data trends and patterns, it is time to train the model with data using various methods and methodologies. This includes the selection and implementation of model techniques, model training, the setup and modification of model hyperparameters, model validation, collecting model development and testing, algorithm selection and model optimisation. Data scientists should choose the appropriate algorithm by considering data needs. In addition, they must determine whether model explainability or interpretability is required, test several model variants, etc. The constructed model can then be evaluated for its functionality.

### Step 6: Model Evaluation

Model approval and assessment during training is a crucial step for determining whether a data scientist has a successful supervised data science model based on numerous metrics. Model planning and assessment is an important stage since it

oversees the selection of a learning strategy or model and provides a measure of the model's performance. Methods such as the ROC curve and cross-validation are utilised to generalise the model output for fresh data. If the model is yielding good findings, data scientists may proceed with its implementation.

### Step 7: Putting Model into Production

This step involves testing the model's performance in the actual world. This stage is also known as the model's "operationalization." Data scientists should deploy the model, continuously monitor its performance and modify various aspects to improve the model's overall performance. Model operationalization might range from merely providing a report to a more complicated, multi-endpoint deployment, depending on the needs of the organisation. Yet, data scientists must assure continual enhancements and iterations, since both technological capabilities and business needs change often.

#### 2.3.4 Presentation and Automation

- **Data Presentation**

Data presentation is the comparison of many data sets using visual aids, such as graphs. With a graph, you may depict the relationship between the information and other data. Following data analysis, this procedure helps organise information by visualising and presenting it in a more comprehensible style. This method is applicable to practically every business since it allows specialists to communicate their results following data analysis.

##### Types Of Data Presentation

You can present data in one of the following three ways:

###### Textual

While presenting data in this manner, you express the link between data using words. Textual presentation helps researchers to convey data that cannot be represented visually. A study's findings are an example of data that may be presented textually. When a researcher want to include extra context or explanation in their presentation, they may opt for this style, as information may display more clearly in text. Textual presentation is typical for communicating research and introducing novel concepts. It contains solely paragraphs and words, with no accompanying tables or graphs.

###### Tabular

Tabular presentation is the dissemination of vast volumes of information via a table. With this strategy, data are arranged in rows and columns based on their qualities. Tabular display facilitates data comparison and information visualisation. This sort of presentation is used in analysis by researchers, such as:

- **Qualitative classification:** country, age, social standing, attractiveness and personality qualities may appear in a table for comparing sociological and psychological data.
- **Quantitative classification:** Items in this category can be counted or numbered.

## Notes

- **Spatial classification:** This pertains to circumstances in which information is based on place, such as city, state, or regional data.
- **Temporal classification:** Time is the variable in this category, thus any measure of time, such as seconds, hours, days, or weeks, can assist in classifying the data.

The advantages of utilising a table to show your data are that it simplifies the data, making it more comprehensible to your audience, helps give a side-by-side comparison of the variables you select and can save space in your presentation by condensing the information.

### Diagrammatic

This technique of data presentation employs diagrams and graphics. It is the most visually appealing style of data presentation and gives a fast overview of statistical data. There are four fundamental types of diagrams:

- **Pictograms:** This diagram represents data with visuals. For instance, to depict the quantity of books sold during the first week of publication, you may depict five books, with each picture representing 1,000 books and 5,000 books purchased by customers.

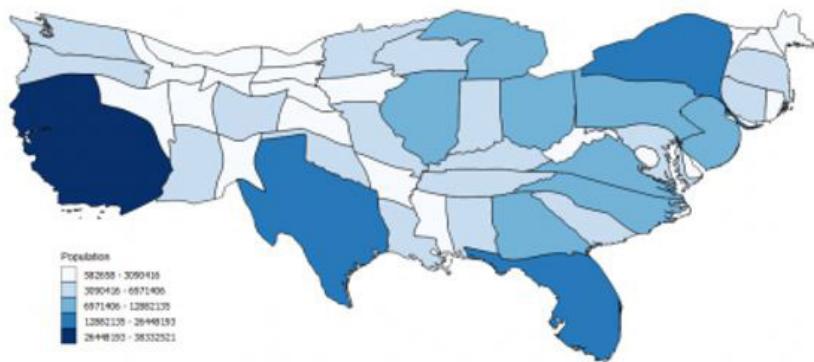
#### Example:

The following pictograph diagram illustrates how many children travelled to school using each method of transportation represented by an image. Each picture in the design indicates a different value.

| Mode of transport | Number of students |
|-------------------|--------------------|
| Bus               | 😊😊😊😊😊😊😊            |
| Car               | 😊😊😊😊               |
| Walking           | 😊😊😊😊😊😊             |
| Bicycle           | 😊😊😊                |

Key: 😊 Represents 3 children

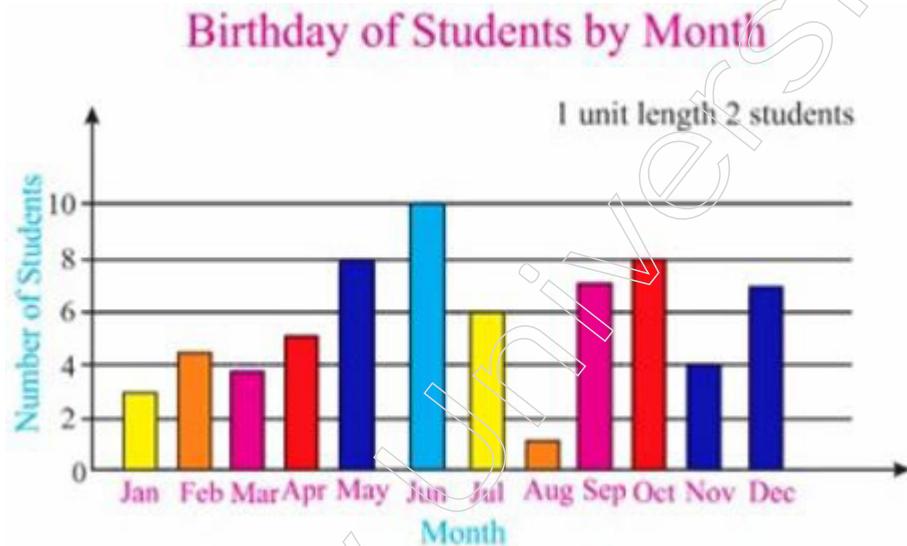
- **Cartograms:** This covers any map depicting the position of a person, location, or thing. For instance, cartograms assist in navigating amusement parks in order to locate attractions, food and gift stores.

**Notes**

- **Bar graphs:** This style employs rectangles of varying widths on the x- and y-axes to depict disparate data values. It displays numerical quantities and data for variables in your research using rectangles.

**Example:**

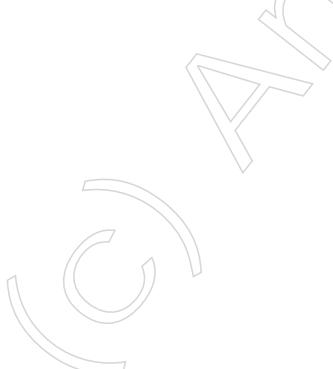
Birthdays of different students at the school in the different months.



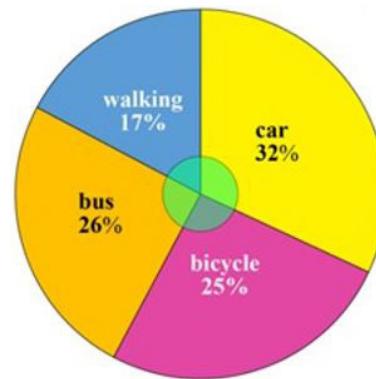
- **Pie Charts:** In this sort of graphic, data is represented as a fraction within a circle. This can display any form of numeric data; however it functions best with fewer variables.

**Example:**

Mode of transport of different students at the school is shown in pie chart below:



## Notes



**Figure: Transport of School**

Diagrams can provide more information about the relationships between variables in the data set than other ways of data presentation because they are more visual. For instance, a bar graph may display data by colour and rectangle size and a more complicated bar graph can be used to display data from numerous variables across time. The diagrammatic design also facilitates rapid data reading and comparison.

- **Data Automation**

Data analytics automation is the examination of digital data using sophisticated computer algorithms and simulations. Depending on the industry in which a company operates, its employees may collect statistical data on consumer information, manufacturing processes, profitability, or performance indicators. Utilizing this data to guide crucial business choices may help a company remain successful, but manually evaluating these data points is time-consuming and expensive.

Automatic analytics solutions save time and money since you can immediately input data into software that creates reports and provides suggestions based on user preferences. This form of automation is particularly valuable for organisations that manage large amounts of data, as there may be several data points to evaluate on a daily basis. By utilising automation software, company owners are able to deliver more dependable outcomes while prioritising other priorities.

### Benefits of Data Analytics Automation

These are some frequent advantages of automating data analytics:

- **Rapid results:** one of the primary advantages of automating data analytics is that algorithms can handle data more quickly than people. By utilising an automated software, you may obtain findings more quickly and spend less time studying individual data points.
- **Handling more data:** In the same amount of time, automation software can filter more data than a team of workers. In addition to being able to process several queries concurrently, data analytics automation tools may analyse larger volumes of user data.
- **Saving money:** Although certain data analytics automation solutions require a paid licence to operate, these programmes can nonetheless save a business money because this endeavour often requires fewer employees and billable hours. These licence costs may be an investment worth making for a firm.

- **Boosting productivity:** As these systems may generate findings more quickly than manual analysis, staff have more time to focus on other crucial responsibilities. Programs that automate data analyses also enable personnel to incorporate freshly reviewed data into project processes, so enhancing the productivity of numerous teams.

## Topic 2.4 Machine Learning

### Introduction

Machine learning (ML) is a sort of artificial intelligence (AI) that enables software programmes to anticipate events with greater precision without being expressly programmed to do so. The input for machine learning algorithms is previous data used to predict future output values. Recommendation engines are a typical use of machine learning. Fraud detection, spam filtering, malware threat detection, business process automation (BPA) and predictive maintenance are further common applications.

Machine learning is essential because it provides businesses with a picture of consumer behaviour trends and company operating patterns and also facilitates the development of new goods. Several of today's biggest corporations, like Facebook, Google and Uber, utilise machine learning extensively. Several businesses now utilise machine learning as a crucial competitive difference.

### 2.4.1 Introduction and Types of Machine Learning

Machine learning is a subfield of artificial intelligence that enables unprogrammed system learning and improvement via experience. Because to its numerous practical uses in a range of sectors, it has become an increasingly popular subject in recent years. Let us cover the fundamentals of machine learning, as well as its application to solving real-world issues. Whether you are a novice hoping to learn about machine learning or a seasoned data scientist seeking to remain abreast of the most recent advancements, we hope you will find something of interest here.

#### What is the definition of Machine Learning?

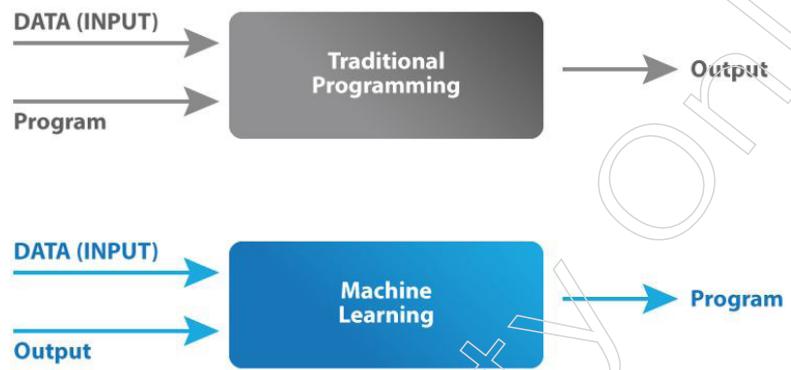
Machine language is a low-level language that computers can understand. It is made up of binary numbers or bits. It is very hard to understand and is also called "machine code" or "object code." Machine language is the only language that a computer can understand. Before being run on a computer, all programmes and programming languages, like Swift and C++, make or run programmes in machine code. Machine language is sent to the system processor whenever a specific job, even the smallest process, is run. Computers are digital machines, so they can only understand binary data. It is founded on the idea that computers can learn from data, see patterns and make decisions with little human input.

This is accomplished with little human interaction, that is, without explicit programming. The process of learning is automated and enhanced depending on the machines' experiences during the process.

Based on high-quality data, several machine learning (ML) model-building techniques are employed to train machines. The method selected relies on the nature of the available data and the task to be automated.

## Notes

You may now be wondering how it differs from conventional programming. In conventional programming, input data and a well-written and verified programme would be fed into a machine to create output. During the learning phase of machine learning, input data and output are provided to the machine and the system figures out a programme on its own. To further comprehend this, please refer to the figure below:



### What are the different types of machine learning?

Classical machine learning is sometimes classed according to how an algorithm learns to make more precise predictions. There are four fundamental learning methodologies: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Data scientists pick an algorithm based on the sort of information they wish to forecast.

- Supervised learning: In this sort of machine learning, data scientists provide algorithms with labelled training data and describe the variables they need the algorithm to evaluate for correlations. The algorithm's input and output are both provided.
- Unsupervised learning: In this sort of machine learning, algorithms are trained on unlabeled data. The system searches through data sets for significant relationships. Both the data used to train algorithms and the predictions or suggestions they provide are predefined.
- Semi-supervised learning: It is a hybrid of the two prior approaches to machine learning. Data scientists may give an algorithm predominantly labelled training data, but the model is allowed to independently explore the data and form its own knowledge of the data set.

Reinforcement learning: Typically, data scientists use reinforcement learning to train a machine to execute a multi-step procedure with precisely stated rules. Data scientists build an algorithm to perform a task and provide it with positive or negative inputs as it determines how to perform the job. Yet for the most part, the algorithm chooses which actions to take along the road on its own.

### 2.4.2 Role of Machine Learning in Data Science

Even though machine learning is always growing with so many new technologies, it is still utilised in a variety of sectors.

Machine learning is essential because it provides businesses with a picture of consumer behaviour trends and operational business patterns and also facilitates the

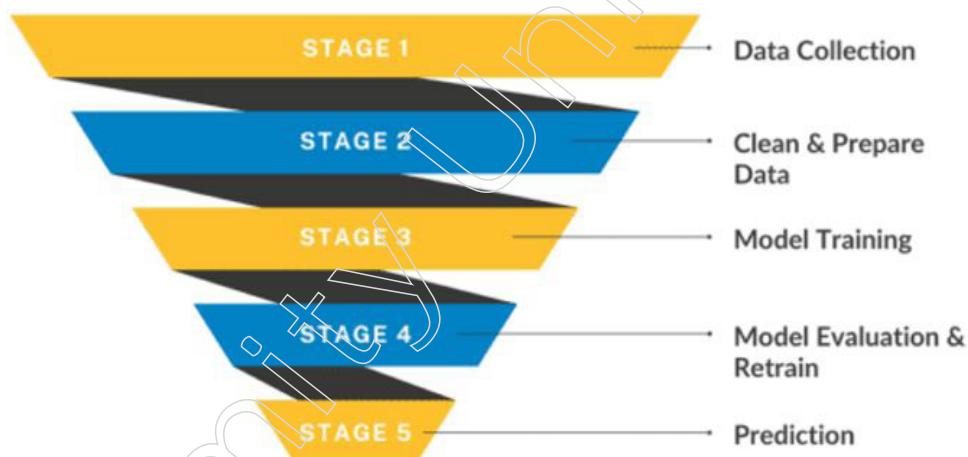
creation of new goods. Several of today's biggest corporations, like Facebook, Google and Uber, utilise machine learning extensively. Several businesses now utilise machine learning as a crucial competitive difference.

Some practical uses of machine learning generate tangible business outcomes, such as time and cost savings, that have the potential to significantly affect the future of your firm. Particularly, machine learning is having a significant impact on the customer service business, helping individuals to do tasks more swiftly and effectively. Machine learning automates, via Virtual Assistant solutions, actions that would normally require a human person to do, such as resetting a password or checking an account balance. This frees up important agent time that may be devoted to the type of customer service that humans execute best: high-touch, complex decision-making that is difficult for a computer to manage. At Interactions, we further enhance the process by eliminating the decision of whether a request should be sent to a human or a machine. Using our proprietary Adaptive Understanding technology, the machine learns its limitations and defers to humans when it lacks confidence in its ability to provide the correct solution.

In data science, we utilise machine learning algorithms when we need to produce accurate estimations about a given collection of data, such as when we need to forecast whether a patient has cancer based on the blood test results. We may do this by providing the algorithm a huge number of cases, including people with and without cancer and their test findings. The system will learn from these instances until it can predict properly if a patient has cancer based on their lab findings.

Hence, machine learning in data science occurs at five stages:

#### Role of Machine Learning in Data Science happens in 5 Stages



##### 1) Data collection

Data gathering is the initial stage of machine learning. According to the business problem, machine learning facilitates the collection and analysis of structured, unstructured and semi-structured data from any database on any system. It may be a CSV file, a PDF, a paper, a picture, or a handwritten form. It is very important to collect accurate and useful data because the quality and amount of data directly affect the results of your Machine Learning Model.

## Notes

### 2) Data preparation and cleansing

In data preparation, machine learning technology facilitates the analysis of data and the creation of features pertinent to the business problem. When properly specified, ML systems comprehend the characteristics and relationships between entities. Features are the foundation of machine learning and every data science effort.

When data preparation is complete, the data must be cleansed, as data in the real world is often contaminated with inconsistencies, noise, incomplete information and missing values. With the use of machine learning, we can locate missing data and perform data imputation, encode categorical columns and eliminate outliers, duplicate rows and null values in an automatic manner.

### 3) Model training

Training a model is dependent on both the quality of the training data and the machine learning technique chosen. A ML method is chosen depending on end-user requirements.

For improved model accuracy, you must also consider model method complexity, performance, interpretability, computer resource needs and speed. After the appropriate machine learning algorithm has been chosen, the training data set is separated into training and testing halves. This is done to determine the model's bias and variance. Model training will result in a functioning model that may be further verified, tested and deployed.

### 4) Model evaluation and retrain

Once model training is completed, there are different metrics to evaluate your model. Note that the selection of a measure is entirely dependent on the model type and implementation strategy. Even if the model has been trained and evaluated, it is not yet prepared to handle your business issues. By refining the parameters of a model, it is possible to increase its precision.

### 5) Model prediction.

While discussing model prediction, it is crucial to comprehend prediction mistakes (bias and variance). Having a thorough grasp of these flaws would allow you to construct correct models and prevent overfitting and underfitting. For a successful data science project, you may limit prediction mistakes further by striking a balance between bias and variance.

In the present day, machine learning (ML) and artificial intelligence (AI) have eclipsed other data science facets in the following ways:

1. Machine learning automatically analyses and investigates vast amounts of data.
2. It automates the process of data analysis and produces predictions in real-time without human intervention.
3. The data model may be further developed and trained to generate real-time predictions. At this phase of the data science lifecycle, machine learning methods are utilised.

### 2.4.3 Classification Algorithms: -Linear Regression, Decision Tree

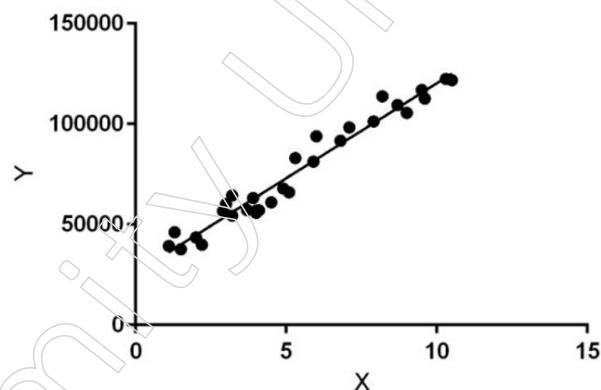
- **Linear Regression**

Linear Regression is a supervised learning-based machine learning technique. It carries out a regression job. Regression models a predicted value based on factors that are independent. Mostly, it is utilised for determining the link between variables and predicting. The type of link considered between dependent and independent variables and the quantity of independent variables distinguish the various regression models. There are several names for the dependent variable in a regression. It is sometimes referred to as an outcome variable, criteria variable, endogenous variable, or regressand. The independent variables are also known as external variables, predictor variables and regressor variables.

Several disciplines, including finance, economics and psychology, employ linear regression to comprehend and forecast the behaviour of a certain variable. For instance, linear regression may be used in finance to determine the link between a company's stock price and its earnings or to forecast the future value of a currency based on its historical performance.

Regression is one of the most important supervised learning tasks. In regression, a series of records with X and Y values are provided and these values are used to develop a function, which may then be used to predict Y from an unknown X. In regression, we must determine the value of Y; hence, a function that predicts Y given continuous XY is necessary.

Hence, Y is referred to as the criteria variable, whereas X is known as the predictor variable. There are a variety of functions and modules that may be utilised for regression. The linear function is the simplest function type. X may indicate a single or numerous characteristics of the situation.



The objective of linear regression is to predict the value of a dependent variable (y) based on a given independent variable (x). Hence, the term is known as Linear Regression. In the image above, X (input) represents job experience and Y (output) represents a person's wage. The regression line provides the best model fit.

#### Hypothesis function for Linear Regression

$$y = \theta_1 + \theta_2 \cdot x$$

During model training, we are given: x: training data input (univariate – a single input variable(parameter)). y: labels to data (Supervised learning) During training the

## Notes

model - the optimal line for predicting the value of  $y$  given a value of  $x$  is fitted. The model gets the best regression fit line by finding the best  $\theta_1$  and  $\theta_2$  values.  $\theta_1$ : intercept  $\theta_2$ : coefficient of  $x$ . Once we find the best  $\theta_1$  and  $\theta_2$  values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of  $y$  for the input value of  $x$ .

### How to update $\theta_1$ and $\theta_2$ values to get the best fit line?

Linear regression is an effective method for comprehending and forecasting the behaviour of a variable, despite its limits. It presupposes a linear connection between independent factors and dependent variables, which is not necessarily true. Moreover, linear regression is sensitive to outliers, or data points that deviate dramatically from the rest of the data. These outliers may have a disproportionate influence on the fitted line, resulting in erroneous predictions.

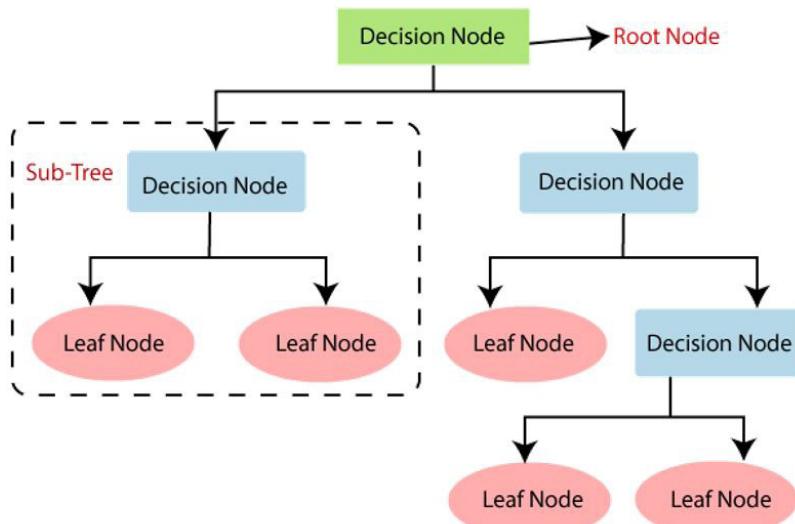
#### ● Decision Tree

##### Decision Tree Classification Algorithm

- A Decision Tree is a technique for Supervised Learning that can be applied to both classification and regression issues, while it is most employed to solve classification problems. It is a classifier with a tree-like structure, where internal nodes represent the characteristics of a dataset, branches represent the decision rules and each leaf node reflects the conclusion.
- A Decision tree contains two nodes: the Decision Node and the Leaf Node. Decision nodes are used to make decisions and have numerous branches, whereas Leaf nodes represent the results of these decisions and do not contain any more branches.
- Decisions or tests are made based on characteristics of the provided dataset.
- It is a graphical depiction for obtaining all potential solutions to a problem/decision based on specified circumstances.
- It is referred to as a decision tree because, similar to a tree, it begins with a root node that branches out and forms a tree-like structure.
- We utilise the CART algorithm, which stands for Classification and Regression Tree algorithm, to construct a tree.
- A decision tree only poses a question and divides the tree into subtrees based on the response (Yes/No).

The figure below describes the structure of a decision tree:

**Note: A decision tree can incorporate categorical data (YES/NO) as well as numeric data.**



### Why use Decision Trees?

The most important consideration when developing a machine learning model is to select the optimal method for the provided dataset and task. Listed below are the two justifications for employing a Decision tree:

- Decision Trees often imitate the way humans make decisions, thus they are simple to comprehend.
- The reasoning underlying the decision tree is easily grasped due to its tree-like form.

### Decision Tree Terminologies

- **Root Node:** The root node is the starting point of the decision tree. It represents the complete dataset, which is then split into two or more homogenous sets.
- **Leaf Node:** Leaf nodes are the last output node and after obtaining a leaf node, the tree cannot be divided further.
- **Splitting:** is the process of separating the decision node/root node into sub-nodes based on the specified requirements.
- **Branch/Sub-Tree:** A tree resulting from a tree's division.
- **Pruning:** Pruning is the removal of unneeded branches from a tree.
- **Parent/Child node:** The root node of a tree is referred to as the parent node, while additional nodes are referred to as child nodes.

### 2.4.4 Naïve Bayes Classifier, K-means

- **Naïve Bayes Classifier**

- The Naïve Bayes algorithm is a supervised learning method that uses Bayes's theorem to solve classification issues.
- It is utilised mostly in text classification tasks that include a high-dimensional training dataset.

## Notes

- The Naïve Bayes Classifier is one of the simplest and most efficient classification algorithms, which aids in the development of fast machine learning models capable of making rapid predictions.
- It is a probabilistic classifier, meaning it makes predictions based on the likelihood of an item.
- Popular applications of the Naïve Bayes Algorithm include spam filtering, sentimental analysis and article classification.

### Why is it called Naïve Bayes?

The Naïve Bayes method consists of the phrases naïve and bayes, which may be defined as follows:

- **Naïve:** It is called Naïve because it implies that the presence of one characteristic is unrelated to the occurrence of other characteristics. For example, if the fruit is classified based on colour, shape and flavour, then fruit that is red, spherical and sweet is labelled as an apple. So, each aspect alone contributes to identifying the object as an apple, without relying on the others.
- **Bayes:** It is referred to as Bayes since it relies on Bayes' Theorem.

### Bayes' Theorem:

- Bayes' theorem, also known as Bayes' Rule or Bayes' law, is used to calculate the probability of a hypothesis based on previous information. It depends on the probabilities under consideration.
- The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

**Where,**

**P(A|B)** is Posterior probability: Probability of hypothesis A on the observed event B.

**P(B|A)** is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.10s

**P(A)** is Prior Probability: Probability of hypothesis before observing the evidence.

**P(B)** is Marginal Probability: Probability of Evidence.

### Working of Naïve Bayes' Classifier:

The following illustration illustrates how the Naïve Bayes Classifier functions. If we have a dataset of weather conditions and a goal variable named "Play". Thus, utilising this dataset, we must select whether or not to play on a given day based on the weather circumstances. So, to resolve this issue, we must do the following steps:

1. Turn the dataset provided into frequency tables.
2. Create the Likelihood table by determining the probability of the characteristics provided.
3. Use Bayes' theorem to determine the posterior probability.

## Where is Naive Bayes Used?

Use Naive Bayes for the following purposes:

- **Facial Recognition**

As a classifier, it is used to identify faces or their other characteristics, such as the nose, mouth, eyes, etc.

- **Weather Forecast**

It may be used to anticipate whether the weather will be favourable or unfavourable.

- **Medical Diagnose**

Physicians can diagnose patients using the information provided by the classifier. Healthcare workers may utilise Naive Bayes to determine if a patient is at high risk for heart disease, cancer and other diseases and disorders.

- **News Classification**

Google News uses a Naive Bayes classifier to determine whether the news is political, international, etc. Because the Naive Bayes Classifier has so many applications, it is beneficial to understand how it operates.

- **K-Means Clustering Algorithm**

K-Means Clustering is an unsupervised learning approach used in machine learning and data science to tackle clustering issues. In this section, we will discuss the K-means clustering technique, how it operates and the Python implementation of k-means clustering.

## What is K-Means Algorithm?

K-Means Clustering is an Unsupervised Learning technique that clusters the unlabelled dataset into several categories. Here K specifies the number of pre-defined clusters that must be produced in the process; if K = 2, there will be two clusters, if K = 3, there will be three clusters, etc.

It is an iterative approach that partitions the unlabeled dataset into k distinct clusters so that each dataset only belongs to one group with identical attributes.

It allows us to cluster the data into several groups and provides a quick method for discovering the categories of groups in an unlabeled dataset without the requirement for training.

It is a centroid-based technique in which each cluster has a corresponding centroid. This algorithm's primary objective is to reduce the total distance between each data point and its matching cluster.

The method receives as input the unlabeled dataset, splits the dataset into k clusters and continues the procedure until the optimal clusters cannot be identified. With this procedure, k should have a preset value.

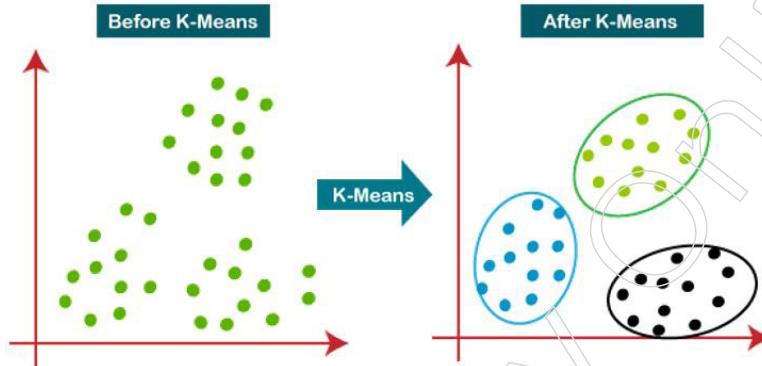
The k-means clustering method primarily accomplishes two functions:

- Iteratively determines the optimal value for K centre points or centroids.
- Assigns each data point to the k-centre nearest to it. The data points that are close to the specific k-centre form a cluster.

## Notes

Hence, each cluster contains datapoints with some commonality and is distinct from the others.

The graphic below illustrates how the K-means Clustering Algorithm operates:



How does the K-Means Algorithm Work?

The K-Means algorithm's operation is outlined in the stages below.

**Step 1:** Determine the number of clusters by selecting K.

**Step 2:** Pick K locations or centroids at random. (It may be different from the input dataset).

**Step 3:** Assign each data point to its nearest centroid, which will construct the K clusters previously determined.

**Step 4:** Compute the variance and assign a new centroid for each cluster in the fourth step.

**Step 5:** Repeat the third steps, which entails reassigning each datapoint to the nearest new cluster centroid.

**Step 6:** Go to step 4 if reassignment happens; otherwise, proceed to FINISH.

**Step 7:** The model is complete.

### 2.4.5 K-Nearest Neighbour, Support Vector Machine

- **K-Nearest Neighbour**

#### K-Nearest Neighbor(KNN) Algorithm for Machine Learning

- K-Nearest Neighbor is one of the most straightforward Machine Learning algorithms based on the Supervised Learning approach.
- The K-NN method considers the similarity between the new case/data and the existing cases and places the new case in the category that is most similar to the existing categories.
- The K-NN algorithm maintains all available data and classifies a new data point on the basis of similarity. This implies that as fresh data becomes available, it may be quickly sorted into a suitable category using the K- NN method.
- The K-NN technique may be used for both Regression and Classification, however it is predominantly employed for Classification tasks.

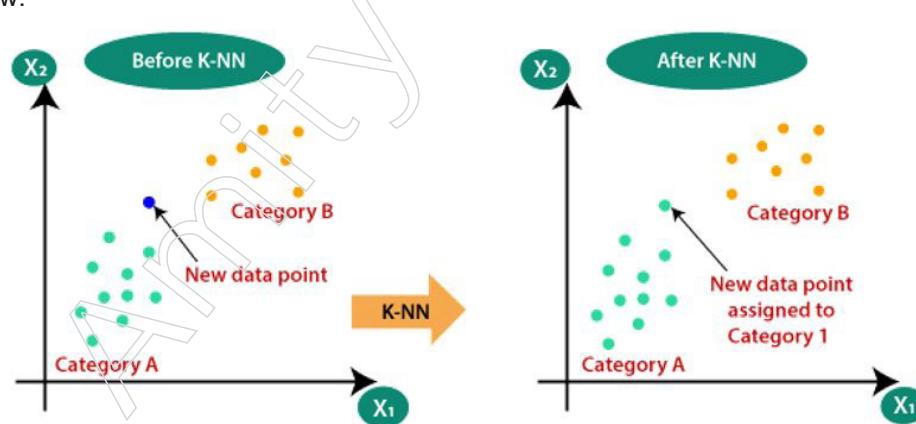
- K-NN is a non-parametric method, hence it makes no assumptions about the underlying data.
- It is also known as a lazy learner algorithm since it does not instantly learn from the training set. Instead, it stores the dataset and takes an action on the dataset at the time of classification.
- During the training phase, the KNN algorithm simply saves the dataset and when it receives new data, it classifies it into a category that is highly similar to the original category.
- Example: Say we have a photograph of a creature that resembles both a cat and a dog, but we want to determine which it is. Thus, we may utilise the KNN method for this identification, as it is based on a measure of similarity. Based on the similarities between the new data set and the photographs of cats and dogs, our KNN model will classify the new data set as either cats or dogs.

## KNN Classifier



### Why do we need a K-NN Algorithm?

If there are two categories, namely Category A and Category B and we obtain a new data point  $x_1$ , we must determine which of these two categories this data point belongs to. To address this sort of issue, a K-NN method is required. We can simply determine the category or class of a certain dataset using K-NN. Consider the diagram below:



### How does K-NN work?

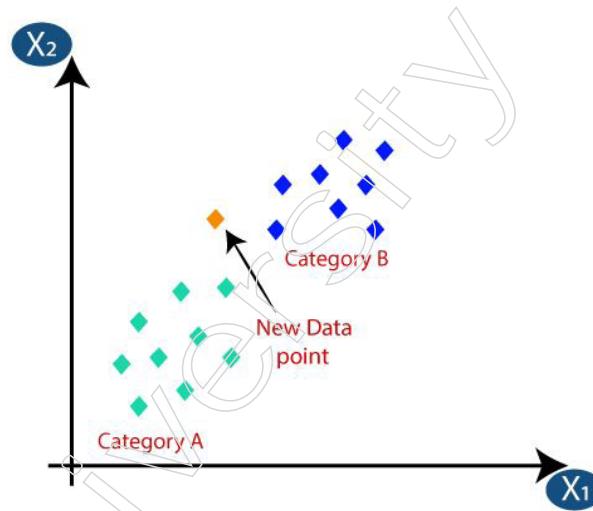
The K-NN working can be explained based on the below algorithm:

- **Step 1:** Choose the number K of the neighbours in this step.

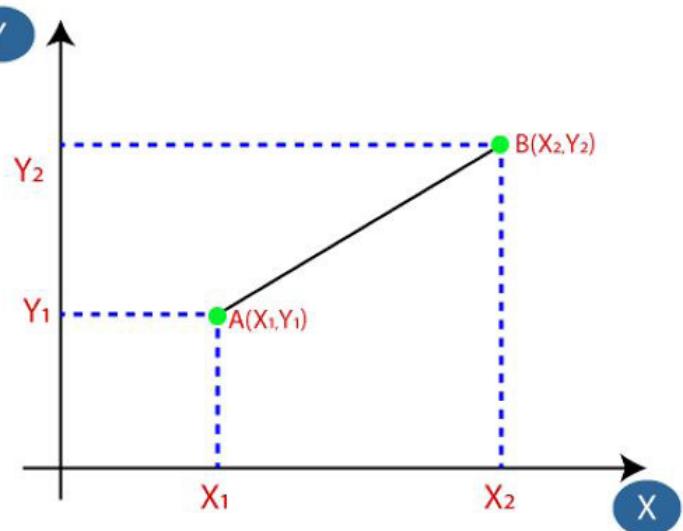
## Notes

- **Step 2:** Compute the Euclidean distance between K neighbours in the second step.
- **Step 3:** Determine the K closest neighbours based on the Euclidean distance.
- **Step 4:** Count the number of data points in each category among these k neighbours.
- **Step 5:** Allocate the new data points to the category with the greatest number of neighbours.
- **Step 6:** Our model is complete.

Assume we have a new data point that has to be assigned to the appropriate category. Consider the image below:

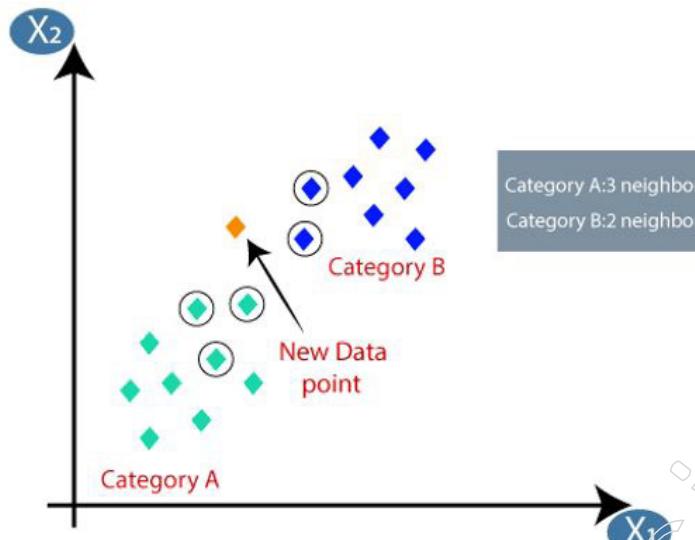


- To begin, we will select the number of neighbours, therefore we will select  $k = 5$ .
- We will next calculate the Euclidean distance between each pair of data points. As previously covered in geometry, the Euclidean distance is the distance between two points. It can be determined by:



$$\text{Euclidean Distance between } A_1 \text{ and } B_2 = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

- By computing the Euclidean distance, we determined the nearest neighbours, which consisted of three neighbours in category A and two in category B. Consider the image below:



- Seeing that the three nearest neighbours all belong to group A, this new data point must also belong to category A.

### How to select the value of K in the K-NN Algorithm?

Following are some considerations to keep in mind while choosing the value of K for the K-NN algorithm:

- There is no specific method for determining the optimal value of "K," thus we must experiment with many values to discover the optimal one. Five is the highest desired value for K.
- An extremely low value for K, such as K=1 or K=2, might be noisy and result in the model exhibiting outlier effects.
- High values for K are desirable, although it may encounter complications.

### Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

### Disadvantages of KNN Algorithm:

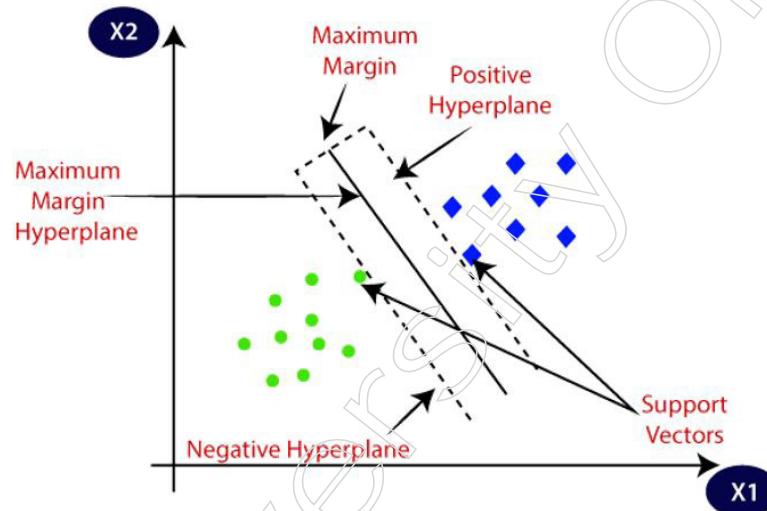
- It is always necessary to ascertain the value of K, which can be difficult at times.
- The calculation cost is considerable as a result of computing the distance between all training sample data points.
- Support Vector Machine

Support Vector Machine, or SVM, is one of the most used techniques for Classification and Regression issues in Supervised Learning. In Machine Learning, it is used mostly for Classification issues.

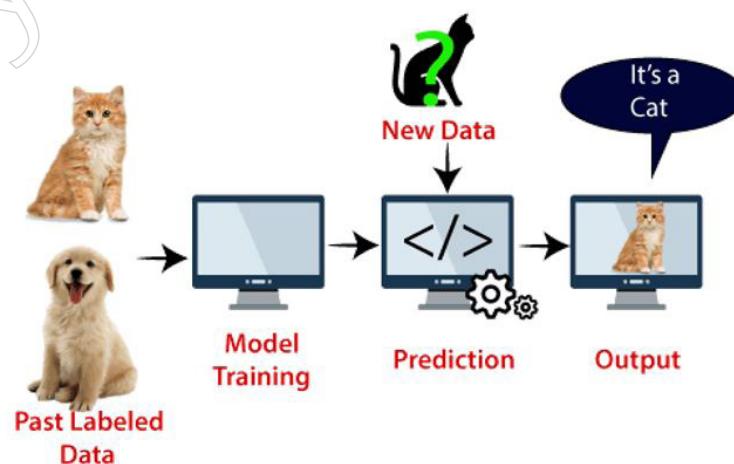
## Notes

The objective of the SVM method is to generate the optimal line or decision boundary that divides n-dimensional space into classes, so that subsequent data points may be readily classified. This optimal decision boundary is referred to as a hyperplane.

SVM selects the extreme points/vectors that contribute to the formation of the hyperplane. These extreme examples are referred to as support vectors and the corresponding technique is known as the Support Vector Machine. Consider the diagram below, which depicts the classification of two distinct categories using a decision boundary or hyperplane:



**Example:** The example provided for the KNN classifier may be utilised to comprehend SVM. The SVM technique may be used to develop a model that can accurately distinguish between a cat and a dog when presented with an unusual cat that also possesses certain canine-like characteristics. First, we will train our model with many photographs of cats and dogs so that it can learn the many characteristics of cats and dogs and then we will test it with this weird species. As a result, as the support vector builds a decision boundary between these two data (cat and dog) and selects extreme instances (support vectors), it will observe the extreme case of cat and dog. According to the support vectors, it will be classified as a cat. Consider the diagram below:



Face identification, picture classification, text categorization, etc. are all possible applications of the SVM method.

### Types of SVM

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which implies that if a dataset can be categorised into two classes using a single straight line, then such data is referred to as linearly separable data and the classifier employed is called Linear SVM.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, meaning that if a dataset cannot be categorised using a straight line, then such data is referred to as non-linear data and the classifier employed is Non-linear SVM.

### Hyperplane and Support Vectors in the SVM algorithm:

Hyperplane: There may be several lines/decision boundaries to separate classes in n-dimensional space, but we must choose the optimal decision boundary for classifying data points. This optimal boundary is known as the SVM hyperplane. The dimensions of the hyperplane are dependent on the features included in the dataset, therefore if there are just two characteristics (as seen in the figure), the hyperplane will be a straight line. Furthermore, if there are three characteristics, hyperplane will be a two-dimensional plane. We always generate a hyperplane with a maximum margin, or maximum distance between data points.

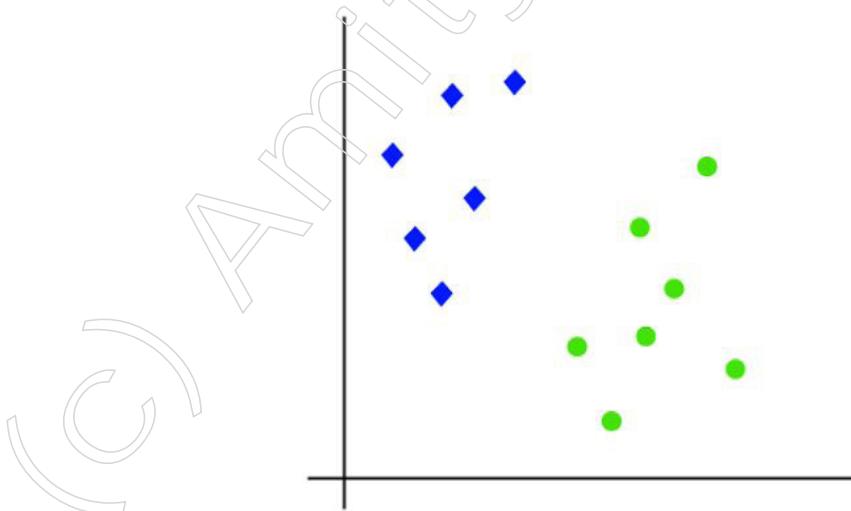
### Support Vectors:

Support Vector refers to the data points or vectors that are closest to the hyperplane and influence the location of the hyperplane. Because these vectors support the hyperplane, they are referred to as Support vectors.

### How does SVM works?

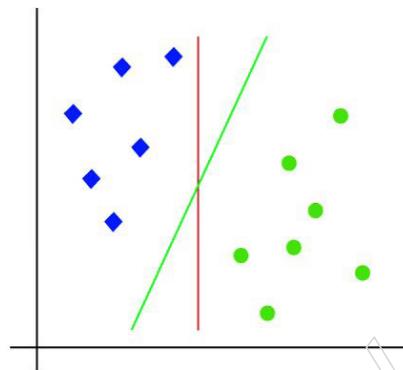
#### Linear SVM:

The operation of the SVM algorithm may be comprehended via the use of an illustration. Assuming we have a dataset with two tags (green and blue) and two features  $x_1$  and  $x_2$  in the dataset. We need a classifier that can categorise the coordinate pair  $(x_1, x_2)$  as either green or blue. Consider the image below:

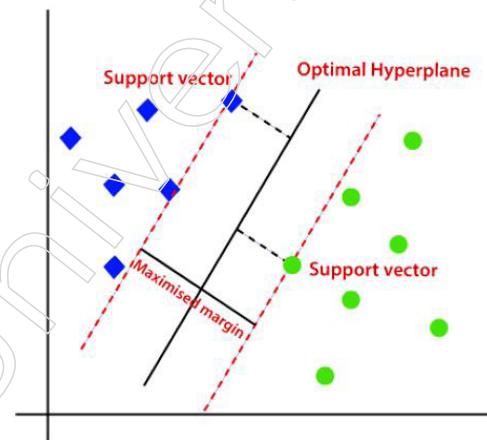


## Notes

Since this is a two-dimensional space, we may simply split these two classes by utilising a straight line. Nonetheless, numerous lines can be used to divide these classes. Consider the image below:

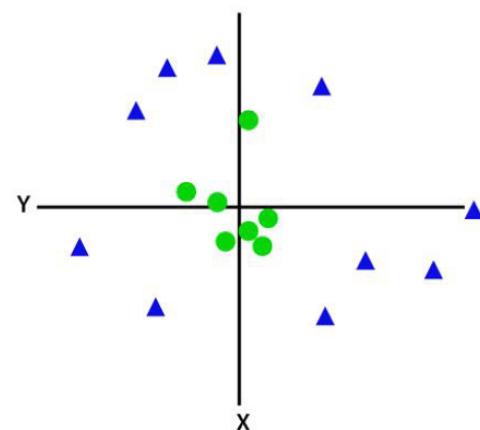


Thus, the SVM method assists in locating the optimal line or decision boundary; this optimal border or region is referred to as a hyperplane. The SVM algorithm identifies the nearest point between two classes' lines. They are known as support vectors. Margin refers to the distance between the vectors and the hyperplane. Therefore, the objective of SVM is to increase this margin. The hyperplane with the greatest margin is known as the ideal hyperplane.



### Non-Linear SVM:

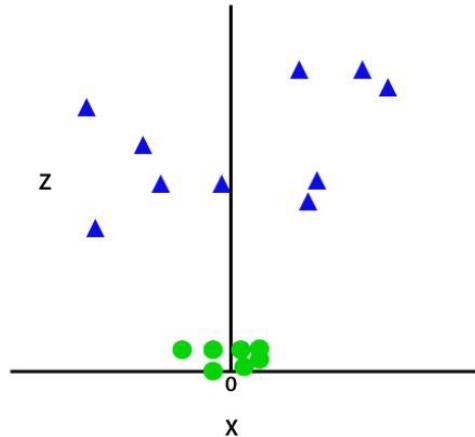
If data is ordered linearly, we can separate it with a straight line; but, for non-linear data, we cannot draw a single straight line. Consider the image below:



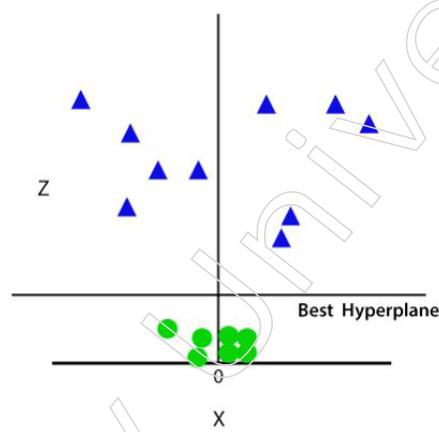
To distinguish these data sets, we must thus add an additional dimension. For linear data, we have utilised dimensions x and y, but for non-linear data, dimension z will be added. It may be expressed as:

$$z=x^2+y^2$$

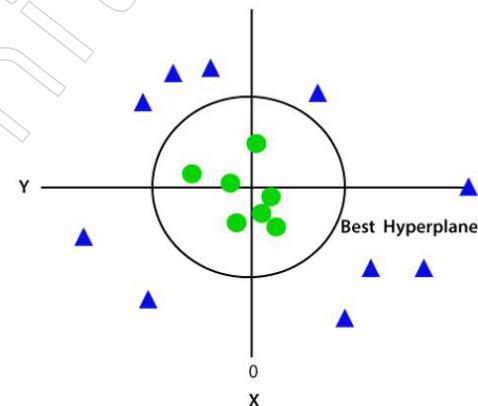
By adding a third dimension, the sample space will resemble the diagram below:



Thus, SVM will split the datasets into classes as follows. Consider the image below:



As we are in 3-D Space, it seems to be a parallel plane to the x-axis. If we transform it to 2D space with  $z = 1$ , the result is:



In the case of non-linear data, we obtain a circumference of radius 1.

## Notes

### Summary

- The term “exploratory data analysis,” or EDA, refers to a technique that is utilised by data scientists to evaluate and study data sets as well as summarise the primary characteristics of such data sets.
- Univariate descriptive statistics focus on analysing just one variable at a time and do not make any comparisons between the variables. Instead, it gives the researcher the opportunity to characterise the factors individually.
- Exploratory data analysis, often known as EDA, is one of the methods utilised in the field of data science with the purpose of identifying key characteristics and trends that are then employed by machine learning and deep learning models.
- Machine learning is a subfield of artificial intelligence that enables unprogrammed system learning and improvement via experience. Because to its numerous practical uses in a range of sectors, it has become an increasingly popular subject in recent years.
- In Supervised learning of machine learning, data scientists provide algorithms with labelled training data and describe the variables they need the algorithm to evaluate for correlations. The algorithm's input and output are both provided.
- In Unsupervised learning of machine learning, algorithms are trained on unlabeled data. The system searches through data sets for significant relationships. Both the data used to train algorithms and the predictions or suggestions they provide are predefined.
- In Semi-supervised learning, which is a hybrid of the two prior approaches to machine learning. Data scientists may give an algorithm predominantly labelled training data, but the model is allowed to independently explore the data and form its own knowledge of the data set.

### Glossary

- **Descriptive statistics:** These are statistics that explain, demonstrate and summarise the fundamental characteristics of a dataset that may be discovered in particular research. These statistics are provided in a summary that summarises the data sample and its measurements.
- **Measures of Central Tendency:** The average or centre of a dataset may be estimated using measures of central tendency and the outcome can be found using one of three methods: the mean, the mode, or the median.
- **Univariate Non-graphical:** This is the simplest type of data analysis since during this type of study, we just look at one variable at a time when researching the data.
- **Central tendency:** The central tendency, also known as the place of distribution, relates to values that are typical or in the centre of the range.
- **Spread:** The spread is a measure of how far out from the centre we are in our search for the information values. Spread may be thought of as a percentage.
- **R:** It is a free and open-source programming language that is used for statistical computation and graphics. R is backed by the R foundation for statistical computing.

- **Variance:** The term variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean (average) and thus from every other number in the set.

### Check Your Understanding

1. What is the full form of EDA?
  - a) Exploratory Data Analysis
  - b) Explanatory Data Analysis
  - c) Exemplary Data Analysis
  - d) Exploratory Division Analysis
2. A \_\_\_\_\_ EDA approach is one that is often used to demonstrate the link between two or more variables using cross-tabulation or statistics.
  - a) Univariate Non-graphical
  - b) Univariate graphical
  - c) Multivariate Non-Graphical
  - d) Multivariate Graphical
3. A \_\_\_\_\_ is a type of bar plot in which each bar reflects either the frequency (count) or the percentage (count divided by total count) of cases for a range of values.
  - a) Leaf Plots
  - b) Histogram
  - c) Quantile Normal Plots
  - d) Stem Plots
4. \_\_\_\_\_ are great for providing information about symmetry and outliers, as well as presenting robust measures of location and spread.
  - a) Scatterplot
  - b) Boxplots
  - c) Heat Map
  - d) Bubble Charts
5. \_\_\_\_\_ is an object-oriented programming language that can be interpreted and has dynamic semantics. Its high level, built-in data structures, in conjunction with dynamic binding, make it a particularly appealing option for the quick creation of applications.
  - a) Python
  - b) R
  - c) Skewness
  - d) OOPS

**Notes**

6. The term \_\_\_\_\_ refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean (average) and thus from every other number in the set.
- Skewness
  - Variance
  - Standard Score
  - Estimated Mean
7. The act of analysing the results and drawing conclusions based on data that has been subjected to random fluctuation is known as \_\_\_\_\_.
- Statistical Inference
  - Standard Deviation
  - Point Estimates
  - Confidence Interval Estimates
8. The definition of \_\_\_\_\_ is the total of all the frequencies that have occurred in the values or intervals that have come before the present one.
- Cumulative Frequency
  - Variance
  - Data Set
  - Score Deviation
9. The term \_\_\_\_\_ refers to the divergence of scores in a group or series from their respective mean values. It is more accurate to say that it relates to the variance of the group's scores in comparison to the mean.
- Variability
  - Visibility
  - Variance
  - Squared Deviations
10. The statistical method known as \_\_\_\_\_ involves putting your presumptions about a population parameter to the test in order to determine whether or not they are accurate.
- Hypothesis Testing
  - Bi-variate regression
  - Multi-variate regression
  - T-Test
11. The \_\_\_\_\_ and the "alternative hypothesis" are the two hypotheses that are tested by every analyst using a population sample that is chosen at random.
- Null hypothesis

**Notes**

- b) Pearson Correlation  
c) Variability of Estimates  
d) Standard Score
12. In the field of statistics, the term \_\_\_\_\_ refers to information that is laid down in the form of a table, complete with rows and columns.
- a) Tabular data  
b) Line Plot  
c) Graph  
d) Pictograph
13. A \_\_\_\_\_, also known as a strip plot or dot chart, is an easy way to visualise data that consists of data points displayed as dots on a graph that has an x- and y-axis.
- a) Scatterplot  
b) Dot Plot  
c) Bar Plot  
d) Tabular Data
14. The circular statistical graphic that is commonly referred to as a \_\_\_\_\_ is also referred to as a “circle chart,” and it illustrates numerical issues by splitting the circle into sectors or portions.
- a) Bar Graph  
b) Line Plot  
c) Pie chart  
d) Tabular Plot
15. The act of representing information via the use of pictures is referred to as \_\_\_\_\_.
- a) Bar Graph  
b) Line Graph  
c) Pictograph  
d) Graph
16. What is the full form of AI?
- a) Artificial Intelligence  
b) Annotation Intelligence  
c) Acute Intelligence  
d) Analytical Intelligence
17. \_\_\_\_\_ is the comparison of many data sets using visual aids, such as graphs. With a graph, you may depict the relationship between the information and other data.

**Notes**

- a) Framing Data
  - b) Data presentation
  - c) Model Building
  - d) Data Retrieval
18. \_\_\_\_\_ technique of data presentation employs diagrams and graphics. It is the most visually appealing style of data presentation and gives a fast overview of statistical data.
- a) Diagrammatic
  - b) Evaluation
  - c) Presentation
  - d) Automation
19. \_\_\_\_\_ is the examination of digital data using sophisticated computer algorithms and simulations.
- a) Data analytics automation
  - b) Machine Learning
  - c) Semi-Supervised Learning
  - d) Data Science
20. What is the full form of ML?
- a) Machine learning
  - b) Machine Lean
  - c) Master Learning
  - d) Managed Learning

**Exercise**

1. Explain the Life cycle and components of data science.
2. Explain the concept of descriptive statistics.
3. Explain the concept of Standard Score.
4. What do you understand by variability of estimates?
5. Explain the concept of data presentation and data automation.

**Learning Activities**

1. Explain the concept of Exploratory Data Analysis.
2. Explain the concept of Pie charts and Graphs.
3. What is Machine Learning and explain its role in data science.

**Check Your Understanding – Answers**

- |       |       |
|-------|-------|
| 1. a) | 2. c) |
| 3. b) | 4. b) |

- |        |        |
|--------|--------|
| 5. a)  | 6. b)  |
| 7. a)  | 8. a)  |
| 9. a)  | 10. a) |
| 11. a) | 12. a) |
| 13. b) | 14. c) |
| 15. c) | 16. a) |
| 17. b) | 18. a) |
| 19. a) | 20. a) |

**Notes****Further Readings and Bibliography**

1. <https://www.geeksforgeeks.org/calculate-the-average-variance-and-standard-deviation-in-r-programming/>
2. <https://www.geeksforgeeks.org/r-statistics/>
3. <https://www.geeksforgeeks.org/mean-median-and-mode-in-r-programming/>
4. <https://www.interaction-design.org/literature/topics/information-visualization>
5. <https://splashbi.com/importance-purpose-benefit-of-data-visualization-tools/>

## Module - III: Feature Selection Algorithms

### Learning Objectives

At the end of this topic, you will be able to understand:

- Analyse extracting feature from data
- Identify transforming features and selecting features
- Identify role of domain expertise
- Describe what is feature selection?
- Define different types of feature selection methods
- Analyse filter methods: types and role
- Describe wrapper method: its different types
- Analyse decision tree: its importance and role in data science
- Describe random forest: its significance

### Introduction

A feature is a property that influences an issue or is relevant for the problem and feature selection is the process of selecting the most significant features for a model. Each phase of machine learning is dependent on feature engineering, which consists mostly of two processes: Feature Selection and Feature Extraction. Even though feature selection and extraction methods may have the same goal, they are quite distinct from one another. Feature selection involves picking a subset of the original collection of features, whereas feature extraction generates new features. Feature selection is a method for minimising the number of model input variables by using only relevant data in order to prevent model overfitting.

### 3.1 Feature Generation

A feature (or column) is a quantifiable piece of data, such as a person's name, age, or gender. It is the fundamental component of a dataset. The quality of a feature can vary considerably and has a substantial impact on the performance of a model. Using methods such as Feature Creation and Feature Selection, we may increase the quality of a dataset's features during pre-processing.

Feature Creation (also known as feature construction, feature extraction and feature engineering) is the process of changing existing features into new, more relevant features. This can be accomplished by mapping a feature into a new feature using a function such as log, or by constructing a new feature from one or several existing features using multiplication or addition.

#### 3.1.1 Extracting Feature from Data

Feature extraction is the process of translating raw data into numerical features that may be handled while keeping the original data set's content. It produces better outcomes than merely applying machine learning to raw data.

Extraction of features can be performed manually or automatically:

- Manual feature extraction entails defining and specifying the important characteristics for a specific situation, as well as designing a method to extract those features. In many circumstances, having a solid grasp of the context or domain can aid in making well-informed judgements on which characteristics may be valuable. Engineers and scientists have created feature extraction algorithms for pictures, signals and text during decades of research. The window mean in a signal is an example of a basic characteristic.
- Automated feature extraction employs specialised algorithms or deep neural networks to automatically extract features from signals or pictures without human interaction. This strategy may be quite beneficial when you want to construct machine learning algorithms rapidly from raw data. Wavelet scattering is an example of the automated extraction of features.

With the rise of deep learning, feature extraction has been superseded by the initial layers of deep neural networks, although mostly for picture data. For signal and time-series applications, feature extraction remains the primary obstacle that necessitates extensive knowledge prior to the development of good prediction models.

The technique of modifying a data collection to better the training of a machine learning model is known as feature engineering. Data scientists expertly modify the training data by adding, removing, merging, or altering variables within the data set to guarantee that the eventual machine learning model will meet its business use case. Data scientists employ feature engineering to generate an input data set that is optimally suited to assist the machine learning algorithm's intended business objective. One way, for instance, includes addressing outliers. Outliers can significantly damage the accuracy of forecasts because they fall so far outside the predicted range. Trimming is a frequent method of dealing with outliers. Trimming only eliminates outlier values, ensuring that they do not contaminate training data.

Extraction of features is a subset of feature engineering. When raw data is inaccessible, data scientists resort to feature extraction. The process of feature extraction converts unprocessed data into numerical features suitable with machine learning algorithms. Raw data in the form of picture files is a frequent application; by extracting the geometry of an item or the redness value from photographs, data scientists may generate new features appropriate for machine learning applications.

The selection of features is strongly connected. Feature selection is the process of deciding which features are most likely to improve the quality of your prediction variable or output, as opposed to feature extraction and feature engineering, which require the creation of new features. Feature selection produces simpler, more readily understood machine learning models by picking just the most pertinent features.

### Feature Extraction Makes Machine Learning More Efficient

Machine learning is made more efficient and accurate via feature extraction. Here are four ways feature extraction helps machine learning models fulfil their intended function more effectively:

- **Reduces redundant data**

Feature extraction eliminates redundant and superfluous data, eliminating the

## Notes

noise. This allows machine learning systems to concentrate on the data that is most pertinent.

- **Improves model accuracy**

The most accurate machine learning models are those created utilising only the data necessary to train the model for its intended business application. Adding peripheral data decreases the model's precision.

- **Boosts speed of learning**

Including training data that does not directly help to the resolution of the business issue slows down the learning process. Models trained on highly relevant data acquire knowledge more rapidly and generate more precise predictions.

- **More-efficient use of compute resources**

Eliminating unnecessary data increases speed and productivity. With fewer data to sort through, fewer computing resources are allocated to tasks that do not generate extra value.

### Feature Extraction Techniques

To exploit the value of raw data sources, data scientists employ several feature extraction techniques. Let us examine three of the most prevalent and how they are utilised to extract machine learning-useful data.

- **Image processing**

Extraction of features serves a crucial function in image processing. This approach is used in conjunction with other tools to recognise characteristics in digital pictures, such as edges, forms and motion. After these have been discovered, the data may be processed to conduct different image analysis-related activities.

- **Bag of words**

This approach, utilised in natural language processing, collects and categorises terms from text-based sources such as web pages, documents and social media postings based on their frequency of occurrence. The bag-of-words approach enables computers to comprehend, analyse and synthesise human language.

- **Autoencoders**

Autoencoders are a sort of unsupervised learning that aims to decrease data noise. Autoencoding involves the compression, encoding and reconstruction of input data. This method uses feature extraction to minimise the dimensionality of data, making it simpler to focus on the input's most essential components.

### Roadblocks To Efficient Feature Extraction

Machine learning is a potent technology, but due to major obstacles, many businesses have yet to utilise it.

- **Poorly designed data pipelines**

Data preparation is one of the most essential steps in machine learning. If poor-quality data is entered, the result will also be of poor quality. Poorly constructed or too

complex data pipelines for machine learning can hinder innovation and are expensive to build and maintain.

- **Compute resource contention**

Programs involving machine learning use large computational resources. Without scalable computational resources, it may be challenging for organisations to allocate the resources necessary to operate a comprehensive machine learning programme while still supporting normal business operations.

- **Siloed data**

Training and deploying machine learning models requires vast volumes of data. Nevertheless, the data of many firms is dispersed across various systems, frequently in different forms. Without a single source of truth, it is impossible to have a comprehensive perspective of the entire organisation.

- **Not fully utilizing AutoML**

As implied by its name, automated machine learning automates a significant portion of the machine learning process. Automatic machine learning (AutoML) expedites activities and reduces the need to manually perform time-consuming processes, allowing machine learning specialists to focus on higher-level responsibilities.

### 3.1.2 Transforming Features

Feature transformation is a mathematical transformation in which we apply a mathematical formula to a specific column (feature) and alter the column's values in a way that is beneficial for further analysis. It is a way for improving the performance of our models. It is also known as Feature Engineering since it produces new features from current features that may improve the performance of the model.

It refers to the class of algorithms that generates new characteristics from existing ones. These new characteristics may not have the same meaning as the original characteristics, but they may have more explanatory power in a different area than in the original space. This is also applicable to Feature Reduction. It can be accomplished in a variety of methods, including via linear combinations of the original characteristics and nonlinear functions. It helps machine learning algorithms to converge faster.

#### Why do we need Feature Transformations?

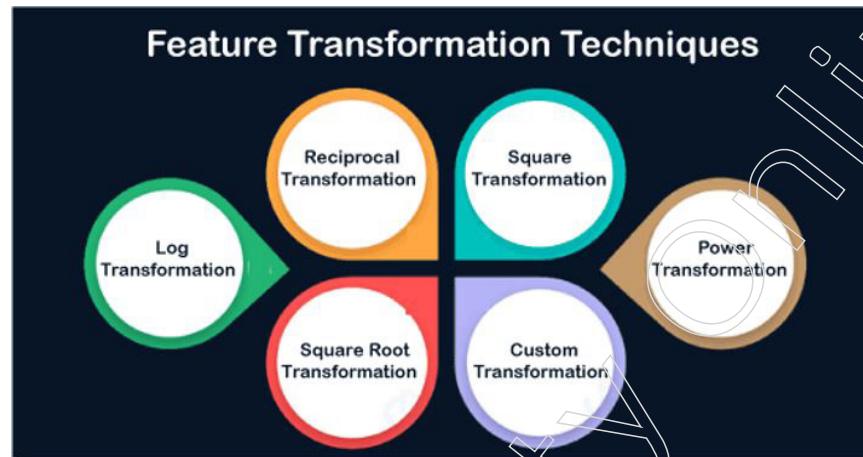
Several data science methods, including Linear and Logistic Regression, assume that the variables follow a normal distribution. Variables in actual datasets are more likely to follow a skewed distribution. By applying various changes to these skewed variables, we may translate this skewed distribution to a normal distribution to improve our models' performance.

As is well-known, the Normal Distribution is a crucial distribution in statistics, used by many statisticians to solve issues. Typically, the distribution of data in nature follows a Normal distribution, such as for age, income, height and weight. Yet, the characteristics of the actual data are not regularly distributed. Still, it is the best estimate when the underlying distribution pattern is unknown.

## Notes

### Feature Transformation Techniques

The following data transformation methods can be used to datasets:



1. **Log Transformation:** In general, these transformations get our data close to a normal distribution, although they cannot adhere to it perfectly. Negative value characteristics are not subject to this change. This transformation is typically done to data that is right-skewed. Transform data from the additive scale to the multiplicative scale, that is, data with a linear distribution.
2. **Reciprocal Transformation:** This transformation is not specified for zero. It is a profound shift with a profound impact. This transformation inverts the order of values of the same sign, making big values smaller and vice versa.
3. **Square Transformation:** This transformation mostly applies to left-skewed data.
4. **Square Root Transformation:** The square root transformation is specified exclusively for positive values. This may be used to reduce the skewness of data that is right-skewed. Log Transformation is superior than this transformation.
5. **Custom Transformation:** A Function Transformer passes its X (and optionally y) inputs to a user-defined function or function object and returns the outcome of this function. If lambda is used as the function, the output transformer will not be pickleable. This is handy for stateless changes such as frequency logarithms, custom scaling, etc.
6. **Power Transformations:** Power transformations are a class of monotonic, parametric changes that make data more Gaussian. Using maximum likelihood, the ideal parameter for stabilising variance and reducing skewness is calculated. This is important for modelling problems with non-constant variance as well as other instances when normalcy is needed. Power Transformer currently supports the Box-Cox and Yeo-Johnson transformations.

Box-cox needs all input data to be positive (even zero is unacceptable), but Yeo-Johnson accepts both positive and negative data. Normalization with zero mean and unit variance is applied by default to the modified data.

- **Box-Cox Transformation:** Sqrt/sqr/log are the special cases of this transformation.
- **Yeo-Johnson Transformation:** It is a variation of the Box-Cox

### 3.1.3 Selecting Features

Before using any approach, it is crucial to comprehend its need, as well as the Feature Selection. To get better results in machine learning, it is required to give a pre-processed and high-quality input dataset. We collect a vast quantity of data to train and improve our model's ability to learn. In general, the dataset includes noisy data, useless data and some helpful data. In addition, the massive volume of data slows down the model's training process and noise and irrelevant data may hinder the model's ability to forecast and perform effectively. In order to exclude these disturbances and unimportant data from the dataset, Feature selection techniques are employed.

Choosing the best characteristics improves the performance of the model. For example, To construct a model that automatically determines which vehicle should be crushed for spare parts, we would need a dataset. This information includes the Model, Year, Owner's name and Mileage for each automobile. Hence, in this dataset, the owner's name does not contribute to the performance of the model because it does not determine whether the automobile should be crushed or not; therefore, we may eliminate this column and choose the remaining features(column) for model creation.

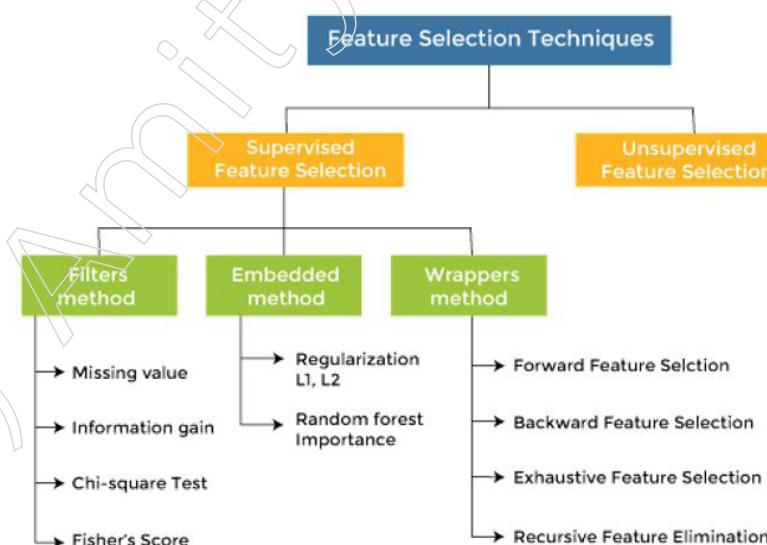
Listed below are a few advantages of feature selection in machine learning:

- This aids in escaping the curse of dimensionality.
- It aids in the simplicity of the model so that researchers may readily comprehend it.
- It decreases training time.
- It decreases overfitting, hence increasing generality.

### Feature Selection Techniques

There are mainly two types of Feature Selection techniques, which are:

- Supervised Feature Selection technique: Techniques for Supervised Feature Selection address the target variable and may be applied to labelled datasets.
- Unsupervised Feature Selection technique: Unsupervised feature selection approaches disregard the target variable and may be used to unlabelled datasets.

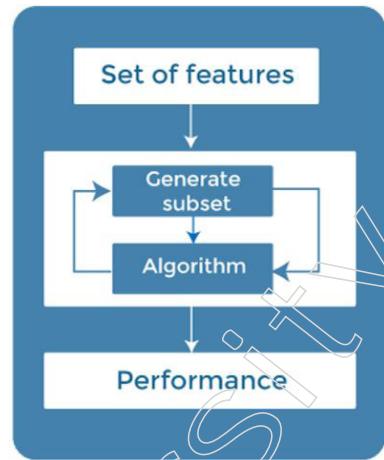


## Notes

Under supervised feature Selection, there are primarily three techniques:

### 1. Wrapper Methods

In wrapper technique, feature selection is approached as a search issue in which several combinations are generated, assessed and compared to other combinations. It trains the algorithm iteratively using the subset of characteristics.



On the basis of the model's output, features are added or deleted and the model is then retrained using the modified feature set. These are examples of wrapper method techniques:

- **Forward selection** - Forward selection is an iterative procedure that starts with an empty collection of features. At each iteration, it adds a new feature and assesses performance to see whether it is enhancing performance. The process is repeated until the inclusion of a new variable or feature no longer improves the model's performance.
- **Backward elimination** - Like forward selection, backward elimination is an iterative method, but it is the inverse of forward selection. This method begins by analysing all of the traits and eliminating the least relevant one. This approach continues until eliminating features no longer improves the model's performance.
- **Exhaustive Feature Selection** - One of the greatest feature selection approaches, exhaustive feature selection assesses each feature set using sheer force. It indicates that this function attempts every conceivable combination of features and returns the best performing collection of characteristics.
- **Recursive Feature Elimination** - Recursive feature elimination is a recursive greedy optimisation technique in which features are picked by recursively selecting a subset of features that is less and smaller. Now, each set of features is used to train an estimator and the relevance of each feature is calculated using `coef_attribute` or through a `feature_importances_attribute`.

### 2. Filter Methods

With the Filter Process, characteristics are chosen based on statistical measurements. This technique is independent of the learning algorithm and selects features as a pre-processing step.

The filter approach eliminates extraneous features and superfluous columns from the model by sorting distinct metrics. Using filter techniques is advantageous since it requires less processing effort and does not overfit the data.

The following are popular Filter Techniques techniques:

- Information Gain
- Chi-square Test
- Fisher's Score
- Missing Value Ratio

**Information Gain:** The information gain determines the decrease in entropy throughout the dataset transformation. It may be used as a strategy for feature selection by computing the information gain of each variable relative to the target variable.

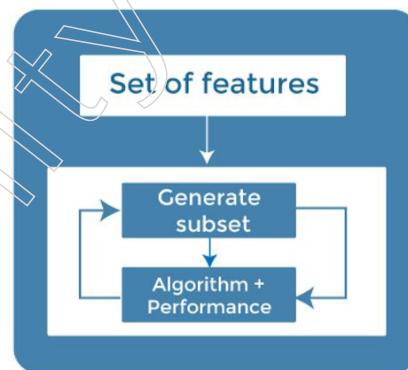
**Chi-square Test:** Chi-square test is a method for determining the association between category variables. Between each feature and the target variable, the chi-square value is computed and the number of features with the highest chi-square value is chosen.

**Fisher's Score:** Fisher's Score is one of the most used supervised techniques for selecting features. It returns the variable's position according to the fisherman's criteria in decreasing order. We may then choose the variables with the highest Fisher's score.

**Missing Value Ratio:** The missing value ratio value may be used to compare the feature set to the threshold value. The missing value ratio is calculated by dividing the number of missing values in each column by the total number of observations. The variable with a value exceeding the threshold can be eliminated.

### 3. Embedded Methods

Embedded techniques integrated the benefits of filter and wrapper methods by taking into account the interaction between features and a low computational cost. These are quick processing techniques comparable to the filter method, but more precise than the filter method.



These approaches are likewise iterative, evaluating each iteration and finding the most significant attributes that contribute the most to training in each iteration. Here are some embedded method techniques:

- **Regularization** - Regularization adds a penalty term to distinct machine learning model parameters to prevent overfitting. The addition of this penalty term to

## Notes

the coefficients reduces some coefficients to zero. Delete the features with 0 coefficients from the dataset. Regularization strategies include L1 Regularization (Lasso Regularization) and Elastic Nets (L1 and L2 regularization).

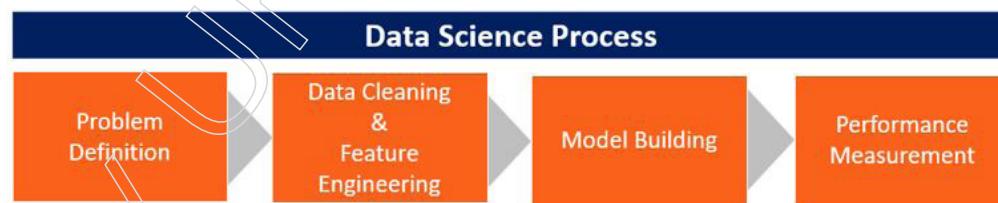
- **Random Forest Importance** - Several tree-based techniques of feature selection give a means of picking features based on feature significance. Here, feature importance describes which feature has a greater influence on the target variable or is of greater significance in model development. Random Forest is a tree-based approach that aggregates a variable number of decision trees using a bagging algorithm. It automatically ranks the nodes based on their performance or decrease in Gini impurity across all trees. The arrangement of nodes according to impurity levels enables the trimming of trees beneath a given node. The remaining nodes constitute a subset of the most vital characteristics.

### 3.1.4 Role of Domain Expertise

Prior to the advent of data science, the phrase "Domain Knowledge" was commonly used. In software engineering, it refers to the knowledge about the operating environment of the target (i.e. software agent). The same notion may be used to data science: "Domain knowledge is knowledge about the context in which data is processed to uncover data secrets." In other words, Domain Knowledge refers to the knowledge about the field to which the data belongs.

### Data Science Process and Domain Knowledge

Here we will explore how domain expertise pertains to each step of the data science process. The data science process may be broken down into the four subprocesses outlined below. The following diagram outlines the process of data science:



#### 1. Problem Definition

Defining the problem to be solved is the first step in any data science. It begins with a general problem description and involves identifying desirable performance criteria. For a basic task, such as forecasting credit default, defining the problem is a straightforward matter of estimating the probability of default based on historical borrower data. Consider, on the other hand, a challenge in robotics or medicine in which a person without domain knowledge cannot even characterise the data pattern they are searching for.

#### 2. Data Cleaning and Feature Engineering

Seldom is the majority of data collected in any field clean and suitable for use. In order to prepare the data for the modelling process, data cleansing and feature engineering are performed. Data cleansing and feature engineering both include data transformation. Incorrectly converted data might result in erroneous conclusions.

For instance, when studying the link between, say, stock price and financial data such as cash flows, one may scale cash flows down. Yet, the scaling would establish a forward-looking bias in the data, as the naive scaling procedure will utilise future data to scale historical data. Any analysis based on improperly converted data will produce erroneous outcomes. In addition, subject expertise is necessary to choose the data attributes that will give the most predictive potential.

### 3. Model Building

In the model-building process, a model is fitted to data. This model is used to solve the problem outlined in the previous stage. The effectiveness of the data science process is dependent on the selection of an acceptable model. Again, this decision relies on the field of application and is facilitated by a solid understanding of the domain.

### 4. Performance Measurement

The last phase of the data science process is performance measurement, which entails assessing how the model performs on fresh or out-of-sample data that was not utilised during model development. The selection of performance measures and thresholds is guided mostly by subject expertise.

When creating a model to forecast credit defaults, for instance, a false negative (predicting a probable defaulter to have good credit) is more expensive than a false positive (predicting a non-defaulter to be a defaulter). These asymmetries would vary throughout fields and it would be difficult to discern them without domain expertise. Costs associated with model failure can only be adequately evaluated by someone with domain expertise.

## Topic 3.2 Feature Selection Algorithms

### 3.2.1 What is Feature Selection?

Feature selection, one of the principal components of feature engineering, is the act of picking the most essential features to feed into machine learning algorithms. Feature selection strategies are used to minimise the number of input variables by removing redundant or unnecessary features and restricting the collection of features to those that are most pertinent to the machine learning model.

The primary advantages of doing feature selection beforehand, as opposed to allowing the machine learning model choose which features are most relevant, are as follows:

- Simpler models: simple models are straightforward to explain; a model that is overly complicated and inexplicable is not valuable.
- Lower training times: a more accurate subset of features reduces the time required to train a model.
- Variance reduction: enhance the precision of estimations produced from a specific simulation.
- Escape the curse of high dimensionality: dimensionally cursed phenomena claims that as dimensionality and the number of features rise, the volume of space expands so rapidly that accessible data become constrained - PCA feature selection may be employed to minimise dimensionality.

## Notes

The most common input variable data types include: Numerical Variables, such as Integer Variables and Floating Point Variables; and Categorical Variables, such as Boolean Variables, Ordinal Variables and Nominal Variables. Popular libraries for feature selection include sklearn feature selection, feature selection Python and feature selection in R.

What makes one variable better than another? Typically, there are three key properties in a feature representation that makes it most desirable: easy to model, works well with regularization strategies and disentangling of causal factors.

### 3.2.2 Different Types of Feature Selection Methods

Feature selection methods are classified as either supervised, for use with labelled data, or unsupervised, for use with unlabelled data. Unsupervised methods can be categorised as filter methods, wrapper methods, embedding methods, or hybrid approaches.

- **Filter methods:** Filter methods choose features based on statistics as opposed to the performance of feature selection cross-validation. The use of a given measure to detect irrelevant qualities and execute recursive feature selection. Filter techniques are either univariate, in which an ordered ranking list of features is created to influence the final selection of feature subset, or multivariate, in which the relevance of the features is evaluated, detecting redundant and irrelevant characteristics.
- **Wrapper methods:** Wrapper feature selection approaches see the selection of a collection of features as a search issue, assessing their quality through the preparation, assessment and comparison of one combination of characteristics to others. This strategy allows the discovery of potential variable interactions. Wrapper approaches concentrate on subsets of features that increase the quality of the clustering algorithm's selection outcomes. Boruta feature selection and Forward feature selection are popular examples.
- **Embedded methods:** Embedded feature selection methods incorporate the feature selection machine learning algorithm as part of the learning process, in which classification and feature selection are conducted concurrently. Carefully extracting the characteristics that will contribute the most to each iteration of model training. Common embedding approaches include random forest feature selection, decision tree feature selection and LASSO feature selection.

### 3.2.3 Filter Methods: Types and Role

#### Filter Methods

Typically, these techniques are employed during the pre-processing phase. These techniques choose characteristics from the dataset regardless of the machine learning algorithm employed. In terms of computing, they are very quick and economical and they are excellent at reducing redundant, correlated and duplicate features, but they do not eliminate multicollinearity. The selection of features is examined separately, which may be advantageous when features are in isolation (don't depend on other characteristics) but will lag when a combination of features might lead to an improvement in the model's overall performance.

Set of all features → Selecting the best subset → Learning algorithm → Performance

## Notes

### Filter Methods Implementation

Some techniques used are:

- **Information Gain** - It is defined as the amount of information a feature provides for identifying the target value and assesses the reduction in entropy values. Information gain is determined for each characteristic based on the goal values for feature selection.
- **Chi-square test** — The Chi-square technique ( $\chi^2$ ) is commonly employed to examine the association between categorical variables. It compares the observed values of the dataset's properties to their predicted value.

$$\chi^2 = \sum \frac{(Observed\ value - Expected\ value)^2}{Expected\ value}$$

### Chi-square Formula

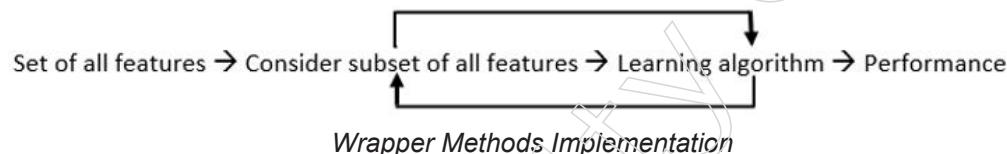
- **Fisher's Score** - Fisher's Score picks each feature individually based on their Fisher criteria scores, resulting in a suboptimal collection of features. The selected attribute is superior the higher Fisher's score.
- **Correlation Coefficient** — Pearson's Correlation Coefficient quantifies the link between two continuous variables and the direction of the relationship, with values ranging from -1 to 1.
- **Variance Threshold** - This method eliminates any characteristics whose variance falls below a specified threshold. This approach eliminates features with zero variance by default. This strategy assumes that traits with more volatility are more likely to hold more information.
- **Mean Absolute Difference (MAD)** – This approach is comparable to the variance threshold method, with the exception of the absence of a square in MAD. This approach computes the average absolute deviation from the mean value.
- **Dispersion Ratio** - Dispersion ratio is defined as the ratio of the Arithmetic mean (AM) to that of Geometric mean (GM) for a given feature. Its value ranges from +1 to  $\infty$  as  $AM \geq GM$  for a given feature. Higher dispersion ratio implies a more relevant feature.
- **Mutual Dependence** - This approach determines whether or not two variables are mutually dependent and offers the quantity of information received for one variable by monitoring the other. Based on the presence/absence of a feature, the quantity of information that feature provides to the prediction of the target is determined.
- **Relief** – This approach examines the quality of characteristics by randomly selecting one instance from the dataset, updating each feature and differentiating between nearby examples based on the distance between the selected instance and the two nearest instances of the same and opposite classes.

## Notes

### 3.2.4 Wrapper Method: Its Different Types

#### Wrapper methods:

Wrapper methods, often known as greedy algorithms, train the algorithm using an iterative subset of characteristics. On the basis of the results drawn from training conducted before to the model's creation, features are added or removed. Typically, the individual training the model defines the stopping conditions for picking the optimal subset, such as when the model's performance degrades or when a certain amount of features is reached. The primary benefit of wrapper methods over filter methods is that they give an optimum collection of features for training the model, resulting in more accuracy than filter methods but at a higher computational cost.



#### Some techniques used are:

- **Forward selection** – This is an iterative procedure in which we begin with an empty collection of features and add the feature that improves our model the most after each iteration. The halting criteria is when the inclusion of a new variable no longer improves the model's performance.
- **Backward elimination** - This method is likewise an iterative strategy in which we begin with all features and delete the least significant feature after each step. The halting criteria is until there is no improvement in the model's performance after removing the feature.
- **Bi-directional elimination** — This methodology use both the forward selection and backward elimination techniques concurrently to arrive at a unique answer.
- **Exhaustive selection** - This technique is considered the brute force method for evaluating subsets of features. It generates all potential subsets, develops a learning method for each subset and then chooses the subset with the greatest model performance.
- **Recursive elimination** - This greedy optimisation technique picks features by recursively evaluating a collection of characteristics that becomes progressively smaller. The estimator is trained using an initial set of features and the importance of those features is determined using feature\_importance\_attribute. The least significant features are then eliminated from the present feature set until the desired number of features remains.

### 3.2.5 Decision Tree: Its Importance and Role in Data Science

A decision tree is a type of supervised machine learning that is used to classify or make predictions based on the answers to a previous set of queries. The model is a form of supervised learning, which means it is trained and evaluated on a dataset containing the intended categorization.

Occasionally, the decision tree may not provide a definitive answer or conclusion. Instead, it may provide options from which the data scientist can make an informed choice. Because decision trees imitate human thought processes, it is typically simple for data scientists to comprehend and interpret the results.

### How Does the Decision Tree Work?

Before we delve into the inner workings of a decision tree, let's define its essential terminology.

- Root node: The starting point of a decision tree.
- Splitting: The division of a node into several sub-nodes.
- Decision node: The point at which a sub-node is divided into additional sub-nodes.
- Leaf node: When a sub-node does not divide further into additional sub-nodes; represents potential outcomes.
- Pruning: The process of removing decision tree subnodes.
- Branch: A subdivision of a decision tree composed of multiple nodes.

A decision tree is similar to a tree. The root node is the foundation of the tree. From the root node emanates a series of decision nodes depicting choices to be made. From the decision nodes emanate leaf nodes that represent the resulting consequences. Each decision node represents a query or branching point and the leaf nodes that emanate from it represent the potential responses. The formation of leaf nodes from decision nodes is analogous to the growth of leaves on a tree branch. This is why each subdivision of a decision tree is called a "branch."

### Types of Decision Trees

There are two primary varieties of decision trees: categorical and continuous. The divisions are determined by the types of outcome variables employed.

- **Categorical Variable Decision Tree**

In a decision tree with categorical variables, the answer neatly falls into one of two categories. Was the outcome of the coin flip heads or tails? Is this creature a reptile or a mammal? In this form of decision tree, data is assigned to a single category based on the decisions made at the tree's nodes.

- **Continuous Variable Decision Tree**

The answer to a continuous variable decision tree is not a simple yes or no. It is also known as a regression tree because the decision or outcome variable is dependent on previous decisions or the type of decision involved.

The advantage of a decision tree with continuous variables is that the outcome can be predicted based on multiple variables, as opposed to a single variable in a decision tree with categorical variables. Predictions are made using decision trees with variables that are continuous. If the appropriate algorithm is selected, the system can be applied to both linear and nonlinear relationships.

## Notes

### Role of Decision Tree

Decision trees continue to be an effective and widespread utility. They are frequently employed by data analysts for predictive analysis (e.g. to develop operations strategies in businesses). In machine learning and artificial intelligence, they are used as training algorithms for supervised learning (i.e. categorising data based on various criteria, such as 'yes' or 'no' classifiers).

In numerous industries, decision trees are used to solve numerous categories of problems. They are utilised in industries ranging from technology and health to financial planning due to their adaptability. Examples include:

- A technology company evaluating expansion possibilities based on an examination of historical sales data.
- A toy manufacturer decides where to spend its limited advertising budget based on demographic data indicating where consumers are likely to purchase.
- Banks and mortgage lenders use historical data to forecast the likelihood of a borrower defaulting on payments.
- Triage in the emergency room could use decision trees to prioritise patient care (based on factors such as age, gender, symptoms, etc.)

### Applications of Decision Trees

#### 1. Assessing prospective growth opportunities

Using historical data, one of the applications of decision trees is to evaluate prospective growth opportunities for businesses. Sales data from the past can be used to create decision trees that may lead to significant changes in a company's expansion and growth strategies.

#### 2. Using demographic data to find prospective clients

Using demographic data to locate prospective customers is a further implementation of decision trees. They can help streamline a marketing budget and make educated decisions regarding the business's target market. In the absence of decision trees, the company's marketing budget may be allocated without a particular demographic in mind, which will have an impact on its overall revenue.

#### 3. Serving as a support tool in several fields

Lenders also use decision trees to predict the likelihood of a customer defaulting on a loan by generating predictive models using the client's historical information. Utilizing a decision tree aids lenders in evaluating a customer's creditworthiness to prevent losses.

Decision trees can also be used in operations research for strategic and logistical planning. They can assist in determining strategies that will assist a company in achieving its objectives. In addition to engineering, education, law, business, healthcare and finance, decision trees can also be utilised in the disciplines of education, law, business and healthcare.

## Advantages of Decision Trees

### 1. Easy to read and interpret

A benefit of decision trees is that their outputs are simple to comprehend and interpret without the need for statistical knowledge. For instance, when using decision trees to present demographic information about customers, marketing department employees can read and interpret the graphical representation of the data without the need for statistical knowledge. The data can also generate essential insights regarding the probabilities, costs and alternatives of the marketing department's numerous strategies.

### 2. Easy to prepare

Compared to other decision-making techniques, decision trees require less data preparation effort. However, users must have readily available data to generate new variables with the ability to predict the objective variable. They can also create data classifications without performing intricate calculations. Users can integrate decision trees with other methods for complex scenarios.

### 3. Less data cleaning required

Once the variables are created, there is less need for data cleansing when using decision trees. On the decision tree's data, absent values and outliers have less significance.

### 3.2.6 Random Forest: Its Significance

Leo Breiman and Adele Cutler patented the Random Forest machine learning algorithm, which combines the output of multiple decision trees to produce a single result. Its adoption has been fueled by its usability and adaptability, as it manages both classification and regression problems.

#### Decision trees

Given that the random forest model is comprised of multiple decision trees, it would be useful to begin by briefly describing the decision tree algorithm. Decision trees begin with a fundamental inquiry, such as "Should I surf?" You can then pose a succession of questions to determine the answer, such as "Is it a long period swell?" and "Is the wind blowing offshore?" These queries constitute the decision nodes in the tree, which serve to partition the data. Each query aids an individual in reaching a conclusion, which is represented by the leaf node. Observations that meet the criteria will take the "Yes" path, while those that do not will take the alternative route. Typically, the Classification and Regression Tree (CART) algorithm is used to train decision trees in order to determine the optimal data subset split. Metrics such as Gini impurity, information gain and mean square error (MSE) can be employed to assess the split's quality. This decision tree illustrates a classification problem with the class labels "surf" and "do not surf."

Although decision trees are prevalent supervised learning algorithms, they are susceptible to bias and overfitting. However, when multiple decision trees form an ensemble in the random forest algorithm, the results are more accurate, especially when the individual trees are uncorrelated.

## Notes

### Ensemble methods

Ensemble learning methods consist of a collection of classifiers, such as decision trees and their predictions are combined to determine the most popular outcome. Bagging, also known as bootstrap aggregation and boosting are the most well-known ensemble methods. In 1996, Leo Breiman ([link resides outside of ibm.com](#)) (PDF, 810 KB) introduced the bagging method; in this method, a random sample of data from a training set is selected with replacement, indicating that the same data point can be selected multiple times. After generating multiple data samples, these models are independently trained and depending on the type of task—regression or classification—the average or preponderance of those predictions result in a more accurate estimate. This method is frequently employed to reduce variance in a chaotic dataset.

### Random forest algorithm

The random forest algorithm is an extension of the bagging method in that it combines bagging and feature randomness to generate a forest of decision trees that are uncorrelated. Feature randomness, also referred to as feature bagging or “the random subspace method” (PDF, 121 KB), generates a random subset of features to ensure minimal correlation between decision trees. This is the most significant distinction between decision trees and random forests. While decision trees consider every conceivable feature division, random forests select only a subset of these features. By taking into account all potential data variability, we can reduce the risk of overfitting, bias and overall variance, resulting in more accurate predictions.

### Benefits and challenges of random forest

When applied to classification or regression problems, the random forest algorithm presents a number of important advantages and challenges. These are just a few:

#### Key Benefits

- **Decreased risk of overfitting:** Decision trees run the risk of overfitting because they tend to closely match all training data samples. When there are a large number of decision trees in a random forest, however, the classifier will not overfit the model because the average of uncorrelated trees reduces the aggregate variance and prediction error.
- **Offers flexibility:** Since random forest can manage both regression and classification tasks with high precision, it is a popular technique among data scientists. Feature bagging also makes the random forest classifier an effective estimation tool for missing values, as it maintains accuracy even when a portion of the data is missing.
- **It is simple to determine the importance of a feature:** Random forest makes it simple to evaluate the contribution or importance of a variable to the model. There are several methods to evaluate the significance of a feature. Typically, Gini importance and mean decrease in impurity (MDI) are employed to determine the degree to which the model's accuracy declines when a particular variable is omitted. Permutation importance, also known as mean decrease accuracy (MDA), is an additional measure of importance. MDA determines the average decrease in accuracy by permuting the feature values in oob samples at random.

### Key Challenges

- **Time-consuming process:** Since random forest algorithms can manage large data sets, they can make more accurate predictions; however, they can be sluggish to process data as they compute data for each individual decision tree.
- **Requires more resources:** Given that random forests process larger data sets, they will require more storage capacity.
- **More Complex:** The prediction of a solitary decision tree is simpler to interpret than that of a forest of decision trees.

### Summary

- A feature is a property that influences an issue or is relevant for the problem and feature selection is the process of selecting the most significant features for a model.
- A feature (or column) is a quantifiable piece of data, such as a person's name, age, or gender. It is the fundamental component of a dataset.
- Feature Creation (also known as feature construction, feature extraction and feature engineering) is the process of changing existing features into new, more relevant features.
- Prior to the advent of data science, the phrase "Domain Knowledge" was commonly used. In software engineering, it refers to the knowledge about the operating environment of the target.
- Recursive feature elimination is a recursive greedy optimisation technique in which features are picked by recursively selecting a subset of features that is less and smaller.
- Before using any approach, it is crucial to comprehend its need, as well as the Feature Selection. To get better results in machine learning, it is required to give a pre-processed and high-quality input dataset.
- Power transformations are a class of monotonic, parametric changes that make data more Gaussian. Using maximum likelihood, the ideal parameter for stabilising variance and reducing skewness is calculated.
- The square root transformation is specified exclusively for positive values. This may be used to reduce the skewness of data that is right-skewed. Log Transformation is superior than this transformation.

### Glossary

- **Feature extraction:** It is the process of translating raw data into numerical features that may be handled while keeping the original data set's content.
- **Regularization** - Regularization adds a penalty term to distinct machine learning model parameters to prevent overfitting. The addition of this penalty term to the coefficients reduces some coefficients to zero.
- **Random Forest Importance** - Several tree-based techniques of feature selection give a means of picking features based on feature significance. Here, feature

## Notes

importance describes which feature has a greater influence on the target variable or is of greater significance in model development.

- **Fisher's Score:** Fisher's Score is one of the most used supervised techniques for selecting features. It returns the variable's position according to the fisherman's criteria in decreasing order. We may then choose the variables with the highest Fisher's score.
- **Missing Value Ratio:** The missing value ratio value may be used to compare the feature set to the threshold value. The missing value ratio is calculated by dividing the number of missing values in each column by the total number of observations. The variable with a value exceeding the threshold can be eliminated.
- **Chi-square Test:** Chi-square test is a method for determining the association between category variables. Between each feature and the target variable, the chi-square value is computed and the number of features with the highest chi-square value is chosen.

### Check Your Understanding

1. \_\_\_\_\_ is the process of translating raw data into numerical features that may be handled while keeping the original data set's content.
  - a) Feature Extraction
  - b) Feature Generation
  - c) Prediction Model
  - d) Feature Restoration
2. The technique of modifying a data collection to better the training of a machine learning model is known as \_\_\_\_\_.
  - a) Feature Engineering
  - b) Resource Contention
  - c) Siloed Data
  - d) Transforming Features
3. \_\_\_\_\_ approach is utilised in natural language processing, collects and categorises terms from text-based sources such as web pages, documents and social media postings based on their frequency of occurrence.
  - a) Bag of words
  - b) Autoencoders
  - c) Image Processing
  - d) Data Pipelines
4. \_\_\_\_\_ are a sort of unsupervised learning that aims to decrease data noise. Autoencoding involves the compression, encoding and reconstruction of input data.
  - a) Autoencoders
  - b) Log Transformation

- c) Reciprocal Transformation  
d) Square Transformation
5. \_\_\_\_\_ is an iterative procedure that starts with an empty collection of features. At each iteration, it adds a new feature and assesses performance to see whether it is enhancing performance.
- a) Forward selection  
b) Custom Transformation  
c) Square Root Transformation  
d) Power Transformation
6. \_\_\_\_\_ is a method for determining the association between category variables. Between each feature and the target variable, the chi-square value is computed and the number of features with the highest chi-square value is chosen.
- a) Box-Cox  
b) Yeo-Johnson  
c) Chi-square test  
d) T-Test
7. \_\_\_\_\_, one of the principal components of feature engineering, is the act of picking the most essential features to feed into machine learning algorithms.
- a) Regularisation  
b) Random Test  
c) Feature selection  
d) Wrapper Method
8. \_\_\_\_\_ is defined as the amount of information a feature provides for identifying the target value and assesses the reduction in entropy values. Information gain is determined for each characteristic based on the goal values for feature selection.
- a) Information Gain  
b) Missing Value  
c) Fisher's Score  
d) Chi-Square Test
9. What is the full form of MAD?
- a) Measure Absolute Difference  
b) Mean Absolute Difference  
c) Missing Absolute Difference  
d) Measure Absolute Division

**Notes**

**Notes**

10. \_\_\_\_\_ is the starting point of a decision tree.
- a) Leaf Node
  - b) Pruning
  - c) Root node
  - d) Subsets
11. \_\_\_\_\_ is the process of removing decision tree subnodes.
- a) Back Node
  - b) Pruning
  - c) Leaf Node
  - d) Subroots
12. \_\_\_\_\_ is when a sub-node does not divide further into additional sub-nodes; represents potential outcomes.
- a) Decision Node
  - b) Splitting
  - c) Leaf node
  - d) Branch
13. The \_\_\_\_\_ is an extension of the bagging method in that it combines bagging and feature randomness to generate a forest of decision trees that are uncorrelated.
- a) Decision Tree
  - b) Learning Algorithm
  - c) Common Predictive Algorithm
  - d) Random Forest Algorithm
14. What is the full form of MDI?
- a) Measure Decrease in Impurity
  - b) Mean Decrease in Impurity
  - c) Mean Demand in Impurity
  - d) Mean Division in Impurity
15. What is the full form of MDA?
- a) Mean Decrease Accuracy
  - b) Measure Decrease Accuracy
  - c) Mean Division Accuracy
  - d) Mean Decrease Automation
16. What is the full form of CART?
- a) Classification and Random Tree

- b) Classification and Reduction Tree  
c) Classification and Regression Tree  
d) Classification and Regression Time
17. What is the full form of MSE?  
a) Mean Square Error  
b) Mean Sequence Error  
c) Mean Simple Error  
d) Mean Square Estimate
18. Wrapper methods, often known as \_\_\_\_\_.  
a) Greedy Algorithms  
b) Linear Regression  
c) Boosting  
d) Multiple Regression
19. \_\_\_\_\_ is a method that eliminates any characteristics whose variance falls below a specified threshold. This approach eliminates features with zero variance by default.  
a) Correlation Coefficient  
b) Mean Absolute Difference  
c) Variance Threshold  
d) Dispersion Ratio
20. \_\_\_\_\_ is knowledge about the context in which data is processed to uncover data secrets.  
a) Fisher's Score  
b) Domain knowledge  
c) Domain Name  
d) Mutual Dependence

**Exercise**

1. Explain the concept of Transforming Features.
2. Explain Features Extraction Techniques.
3. Explain Wrapper methods and its types.

**Learning Activities**

1. Explain the concept of Feature Extraction.
2. Explain the concept of Feature Selection.

**Notes****Check Your Understanding – Answers**

1. a)
2. a)
3. a)
4. a)
5. a)
6. c)
7. c)
8. a)
9. b)
10. c)
11. b)
12. c)
13. d)
14. a)
15. a)
16. c)
17. a)
18. a)
19. c)
20. b)

**Further Readings and Bibliography**

1. <https://splashbi.com/importance-purpose-benefit-of-data-visualization-tools/>
2. <https://www.geeksforgeeks.org/what-is-data-visualization-and-why-is-it-important/>
3. <https://www.tibco.com/reference-center/what-is-information-visualization#:~:text=Information%20visualization%20is%20the%20practice,digestible%20format%20for%20non%2Dexperts.>
4. <https://clauswilke.com/dataviz/aesthetic-mapping.html>

# Module -IV: Recommendation Systems

Notes

## Learning Objectives

At the end of this topic, you will be able to understand:

- Analyse what is predictive modelling?
- Identify what is dimensionality reduction?
- Describe importance of dimensionality reduction
- Analyse different components of dimensionality reduction
- Identify need for dimensionality reduction
- Describe understanding singular value reduction: mathematical concept
- Learn about single value theorem
- Identify what is PCA?
- Analyse algorithm for PCA
- Identify application and role of PCA in dimensionality reduction
- Describe example for finding PCA in dataset

## Introduction

Dimensionality refers to the number of input features, variables, or columns present in a given dataset and dimensionality reduction refers to the process of reducing these features. A dataset contains a large number of input features in various circumstances, which complicates the task of predictive modelling. In situations where a large number of features in the training dataset makes it challenging to visualise or make predictions, dimensionality reduction techniques are required.

### 4.1 Dimensionality Reduction

Dimensionality reduction technique can be defined as “the process of transforming a dataset with greater dimensions into a dataset with fewer dimensions while preserving the same information.” These techniques are extensively employed in machine learning to obtain a more accurate predictive model when solving classification and regression issues. It is frequently employed in disciplines that deal with high-dimensional data, such as speech recognition, signal processing, bioinformatics, etc. It can also be used for data visualisation, noise reduction and cluster analysis, among other applications.

#### 4.1.1 What is Predictive Modelling?

Predictive modelling is a probabilistic method for forecasting outcomes based on a set of predictors. These predictors are essentially characteristics that play a role in determining the model's ultimate outcome.

Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while retaining the maximum amount of information. This can be done for a variety of reasons, including reducing the intricacy of a model, enhancing

## Notes

the performance of a learning algorithm, or making the data simpler to visualise. Dimensionality reduction techniques include principal component analysis (PCA), singular value decomposition (SVD) and linear discriminant analysis (LDA) (LDA). Each technique uses a unique method to project the data onto a lower-dimensional space while preserving the most essential information.

### Types of Predictive Models

Fortunately, prognostic models do not need to be developed from inception for every application. The models and algorithms utilised by predictive analytics tools can be applied to a wide range of use cases. Over time, predictive modelling techniques have been refined. We are able to do more with these models as we add more data, more powerful computing, AI and machine learning and as analytics as a field advances.

#### Predictive analytics models are:

1. **Classification model:** Considered the simplest model, the classification model categorises data for straightforward query responses. To answer the query "Is this a fraudulent transaction?" is an example use case.
2. **Clustering model:** This model groups data based on shared attributes. It functions by clustering objects or people with similar characteristics or behaviours and planning strategies on a larger scale for each group. An example is determining a loan applicant's credit risk based on the past actions of individuals in the same or a similar situation.
3. **Forecast model:** This is a popular model that works on anything with a numeric value and is based on historical data learning. For instance, when determining how much lettuce a restaurant should order for the upcoming week or how many contacts a customer service agent should be able to manage per day or week, the system refers to historical data.
4. **Outliers model:** This model operates by analysing data elements that are abnormal or outliers. For instance, a bank may use an outlier model to detect fraud by determining if a transaction deviates from a customer's normal purchasing patterns or if an expense in a given category is unusual. For instance, a \$1,000 credit card charge for a washer and dryer at the cardholder's favoured large box store would not be cause for alarm. However, a \$1,000 credit card charge for designer clothing at a store where the customer has never charged other items could indicate a compromised account.
5. **Time series model:** This model evaluates a sequence of data points based on the passage of time. For instance, the number of stroke patients admitted to the hospital over the past four months is used to estimate the number of patients the hospital can expect to admit next week, next month and for the remainder of the year. A singular metric that is measured and contrasted over time is therefore more informative than an average.

### Techniques

The following techniques are used in predictive modeling:

1. **Linear Regression:** When two continuous variables exhibit a linear relationship, a linear regression can be used to determine the value of the dependent variable based on the independent variable.
2. **Multiple Regression:** Similar to linear regression, with the difference that the value of the dependent variable is determined by analysing multiple independent variables.
3. **Logistic regression:** It is used to determine dependent variables when the data set is large and categorization is required.
4. **Decision Tree:** It is a commonly used data mining technique. The formulation of a flowchart representing an inverted tree. Here, the internal node divides into branches that enumerate two or more potential decisions and each decision is further subdivided to illustrate additional potential outcomes. This method facilitates the selection of the finest option.
5. **Random Forest:** It is a prominent model for regression and classification. It is used to solve algorithms for machine learning. It consists of distinct decision trees that are unrelated to one another. Collectively, these decision trees facilitate the analysis.
6. **Boosting:** As its name suggests, boosting facilitates learning from the results of other models, such as decision tree, logistic regression, neural network and support vector machine.
7. **Neural Networks:** It is a problem-solving mechanism utilised in machine learning and artificial intelligence. It creates a set of algorithms for a system of computational learning. The three layers of these algorithms are input, processing and output.

### Common Predictive Algorithms

One of machine learning or deep learning is utilised by predictive algorithms. Each constitutes a subset of artificial intelligence (AI). Machine learning (ML) requires structured data such as spreadsheets and machine data. Deep learning (DL) deals with unstructured data such as video, audio, text, social media posts and images — essentially the information humans use to communicate that is not a number or a metric.

Among the most prevalent predictive algorithms are:

1. **Random Forest:** This algorithm is derived from a collection of unrelated decision trees and can classify enormous quantities of data using both classification and regression.
2. **Generalized Linear Model (GLM) for Two Values:** This algorithm reduces the list of variables to determine the “greatest suit.” It can calculate inflection points and alter data acquisition and other influences, such as categorical predictors, to determine the “best fit” outcome, thereby overcoming the limitations of other models, such as linear regression.
3. **Gradient Boosted Model:** Like Random Forest, this algorithm combines multiple decision trees, but unlike Random Forest, these trees are related. It constructs one tree at a time, allowing the subsequent tree to rectify defects in the preceding tree. It is frequently employed in evaluations, such as search engine results.

## Notes

4. **K-Means:** K-Means, a popular and quick algorithm, clusters data elements based on their similarities and is therefore frequently used for clustering models.
5. **Prophet:** This algorithm is utilised in time-series or forecast models for capacity planning, including inventory requirements, sales quotas and resource allocation. It is highly adaptable and can accommodate heuristics and a variety of useful assumptions with ease.

### Limitations of Predictive Modeling

According to a report by McKinsey, common limitations and their “recommended solutions” are as follows:

1. **Errors in data labeling:** These can be circumvented by reinforcement learning or generative adversarial networks (GANs).
2. **Shortage of massive data sets needed to train machine learning:** Shortage of vast data sets required to train machine learning: “one-shot learning,” in which a machine learns from a limited number of demonstrations as opposed to a massive data set, is a potential solution.
3. **The machine’s inability to explain what and why it did what it did:** Machines do not “think” or “learn” like humans do. Similarly, their computations can be so exceptionally complex that humans have difficulty locating the logic, let alone following it. All of this makes it difficult for machines and humans to articulate their work. Nonetheless, model transparency is necessary for a variety of reasons, human safety being the most important. Local-interpretable-model-agnostic explanations (LIME) and attention techniques are promising potential solutions.
4. **Generalizability of learning, or rather lack thereof:** Unlike humans, machines struggle to apply what they have learned. In other words, they have difficulty applying what they have learned to new situations. Only one use case is pertinent to whatever it has learned. This is primarily why we do not need to be concerned about the imminent rise of AI overlords. For machine learning-based predictive modelling to be reusable, that is, applicable to more than one use case, transfer learning is a potential solution.
5. **Bias in data and algorithms:** Non-representation can distort outcomes and result in the maltreatment of vast human populations. Additionally, baked-in biases are difficult to detect and eliminate later. In other words, biases tend to perpetuate themselves. This is a fluid target for which no definitive solution has yet been identified.

#### 4.1.2 What is Dimensionality Reduction?

Dimensionality reduction is the process of reducing the number of variables in a training dataset used to build machine learning models. The process monitors the dimensionality of data by mapping high-dimensional data to a lower-dimensional space that encapsulates the “essence at its core” of the data. The analysis of data with millions of features requires multiple sources and computations. In addition, it requires considerable manual labour. Dimensionality reduction simplifies this complex task by converting a high-dimensional dataset to a lower-dimensional dataset without altering the original dataset’s essential properties. Before beginning the training cycle of

machine learning models, this procedure exposes the data pre-processing stages that are executed.

### Benefits Of Dimensionality Reduction

Dimension reduction is useful for AI engineers and data professionals working with massive datasets, visualising and analysing complex data.

1. It facilitates in data compression, resulting in the need for less storage space.
2. It expedites the calculation.
3. Additionally, it helps remove any unnecessary features.

### Disadvantages Of Dimensionality Reduction

1. We lost some data during the process of dimensionality reduction, which may affect the performance of future training algorithms.
2. It may require a great deal of processing capacity.
3. The interpretation of transformed characteristics may be difficult.
4. Therefore, the independent variables become more challenging to understand.

### Dimensionality Reduction in Predictive Modeling

Dimensionality reduction can be illustrated using a straightforward email classification problem in which we must determine whether or not an email is spam. This may include a variety of characteristics, such as whether the email uses a template, its content, whether it has a generic subject line, etc.

Nonetheless, some of these characteristics may overlap. In another instance, a classification issue that depends on rainfall and humidity can be reduced to a single underlying characteristic due to their strong correlation. Consequently, we may reduce the amount of features in these issues. In contrast to 2-D and 1-D problems, which can be translated to a simple 2-dimensional space, it may be difficult to visualise a 3-D classification problem. In the image below, a 3-D feature space is divided into two 2-D feature spaces to illustrate this concept. If it is subsequently determined that the two feature spaces are related, additional feature reduction may be feasible.

### Dimensionality Reduction Methods and Approaches

Now that we know the extent to which dimensionality reduction contributes to machine learning, the question becomes: what is the most effective way to carry out this process? We have provided a list of the primary techniques that you may take, further subdividing them into a variety of other ways. These strategies and procedures are collectively referred to as Dimensionality Reduction Algorithms in some circles.

#### Feature Selection

The process of picking the optimal and relevant characteristics from an input data collection and deleting features that are not relevant is known as feature selection.

- Filter methods. A useful subset of the data set is obtained through the application of this strategy.

## Notes

- Wrapper methods. The performance of the characteristics that are put into this model is evaluated using this technique, which makes use of the machine learning model. It is up to the performance to decide if it is preferable to maintain or get rid of the characteristics in order to increase the accuracy of the model. This technique provides more precise results than filtering, but at the expense of increased complexity.
- Embedded techniques. The machine learning model's numerous training rounds are inspected by the embedded process, which also determines the relative significance of each feature.

### a) Feature Extraction

By this procedure, the space that possesses an excessive number of dimensions is converted into a space that possesses less dimensions. This procedure is helpful for preserving all of the information while reducing the amount of resources used during the processing of the information. The following are the three methods of extraction that are utilised most frequently.

- Linear discriminant analysis. Dimensionality reduction in continuous data is a typical use of the LDA technique. The data are rotated and projected in the direction of having a higher variance while using LDA. The characteristics that exhibit the greatest amount of variation are referred to as the primary components.
- Kernel PCA. This method is a nonlinear extension of principal component analysis (PCA) that is applicable to more complex structures that cannot be represented in a linear subspace in an easy or suitable manner. It is useful for analysing data. In order to create nonlinear mappings, KPCA makes advantage of the “kernel technique.”
- Quadratic discriminant analysis. This method projects the data in a way that achieves the highest possible level of class separability. Examples belonging to the same category are clustered together in the projection, whilst examples belonging to separate categories are spaced further apart.

### Dimensionality Reduction Techniques

Here are some techniques machine learning professionals use.

#### • Principal Component Analysis

Principal component analysis, or PCA, is a method for reducing the number of dimensions in large data sets by consolidating a large collection of variables into a smaller set that retains the majority of the information contained in the larger set. Since machine learning algorithms can analyse data much more swiftly and efficiently with smaller data sets because there are fewer unnecessary factors to evaluate, accuracy must undoubtedly decrease as the number of variables in a data set decreases. Nevertheless, the solution to dimensionality reduction is to sacrifice some precision for simplicity. PCA endeavours to preserve as much information as possible while minimising the number of variables in a data set.

#### • Backward Feature Elimination

Backward elimination improves the model's performance by beginning with all of its

characteristics and gradually removing the least significant one. We will continue until we observe no improvement when removing features.

1. Initially, all model variables should be employed.
2. Eliminate the variable with the least value (based on, for example, the lowest loss in model accuracy), then continue until a predetermined set of requirements is met.

- **Forward Feature Selection**

Forward selection is an iterative procedure that begins with no features in the dataset. Each iteration introduces new features to improve the model's functionality. Functionality is preserved if efficacy is enhanced. Features that do not improve the results are eliminated. The procedure is repeated until the model no longer improves.

- **Missing Value Ratio**

Think about receiving a dataset. Which arrives first? Obviously, you would need to examine the data prior to constructing a model. As you investigate your data, you realise that it contains absent values. What follows? You will investigate the source of these missing values prior to attempting to impute them or eradicating the variables with missing values entirely.

What if there are too many missing data, say more than 50 percent? Should the variable be removed, or should the absent values be substituted? Given that the variable won't contain many values, we should eliminate it. This is not a certainty, however. We may establish a threshold number and if the proportion of missing data for any variable exceeds that level, we will need to eliminate the variable.

- **Low Variance Filter**

Similar to the Missing Value Ratio method, the Low Variance Filter employs a threshold. However, testing data columns in this instance. The method calculates each variable's variance. All data columns with variances below the threshold are eliminated, as low variance characteristics have no effect on the target variable.

- **High Correlation Filter**

This procedure employs two variables containing identical information, thereby potentially degrading the model. Using the Variance Inflation Factor (VIF), we identify the variables with a high correlation and then select one. Variables with a higher value ( $VIF > 5$ ) can be removed.

- **Decision Trees**

Decision trees are a prevalent algorithm for supervised learning that divides data into homogeneous categories based on input variables. This method addresses issues such as data outliers, absent values and the identification of significant variables.

- **Random Forest**

This method is similar to the decision tree technique. In contrast, we generate a large number of trees (hence "forest") against the target variable in this instance. Then, we identify subsets of features using the usage statistics for each attribute.

- **Factor Analysis**

Imagine there are two variables: education and income. Given that people with

## Notes

higher levels of education tend to have substantially higher incomes, there may be a strong correlation between these variables. The Factor Analysis method categorises variables based on their correlations; therefore, all variables in one category will have a strong correlation among themselves but a feeble relationship with factors in another category (s). Each cohort is referred to as a factor in this context. These variables are few compared to the original dimensions of the data. These elements are difficult to observe, however.

### 4.1.3 Importance of Dimensionality Reduction

Dimensionality reduction has many benefits for machine learning data, including:

- Fewer features means less complexity.
- You will require less storage space because you have fewer data.
- Fewer features require less computation time.
- Model accuracy improves as a result of fewer misleading data.
- Algorithms train quicker as a result of fewer data.
- Reducing the data set's feature dimensions enables speedier data visualisation.
- It eliminates noise and redundant features.
- It reduces the amount of time and storage space needed.
- It helps remove multi-collinearity, which enhances the interpretation of the machine learning model's parameters.
- It is simpler to visualise the data when the dimensions are reduced to 2D or 3D.
- It circumvents the curse of dimensionality.
- It eliminates irrelevant features from the data, as having irrelevant features in the data can reduce the accuracy of the models and cause them to learn based on irrelevant features.

### 4.1.4 Different Components of Dimensionality Reduction

There are two components of dimensionality reduction:

#### • Feature selection

In machine learning, the objective of feature selection techniques is to identify the optimal set of features that enables the development of optimised models of studied phenomena. In machine learning, the techniques for feature selection can be broadly classified into the following categories:

**Supervised Techniques:** These techniques can be applied to labelled data in order to identify the most important features for improving the performance of supervised models such as classification and regression. Example: linear regression, decision tree, support vector machine, etc.

**Unsupervised Methods:** These methods can be applied to unlabeled data. Examples include K-Means Clustering, Principal Component Analysis and Hierarchical Clustering. Here, we must identify a subset of the initial set of variables. In addition, we

need a subset to model the problem. Typically, there are three steps:

- **Filter**

In wrapper methodology, feature selection is approached as a search problem in which various combinations are generated, evaluated and compared to other combinations. It trains the algorithm iteratively using the subset of features.

- **Wrapper**

In the Filter Method, features are chosen based on statistical measurements. This technique is independent of the learning algorithm and selects features as a pre-processing phase. The filter method eliminates irrelevant features and redundant columns from the model by rating distinct metrics. Utilizing filter methods is advantageous because it requires little computational time and does not overfit the data.

- **Embedded**

Embedded methods incorporate the benefits of filter and wrapper methods by taking into account the interaction between features and a low computational cost. These are quick processing techniques comparable to the filter method, but more precise than the filter method. These methods are also iterative, evaluating each iteration and finding the most essential features that contribute the most to training in each iteration.

### b. Feature Extraction

This reduces the data from a high-dimensional space to a lower-dimensional space, or a space with fewer dimensions. Feature extraction is the process of transforming unprocessed data into numerical features that can be processed while preserving the original data set's information. It produces superior results than explicitly applying machine learning to raw data.

Extraction of features can be performed manually or automatically:

- Manual feature extraction requires identifying and describing the pertinent features for a specific problem, as well as implementing a method to extract those features. In many situations, having a solid grasp of the context or domain can aid in making well-informed decisions regarding which features may be beneficial. Engineers and scientists have devised feature extraction methods for images, signals and text over decades of research. The mean of a window in a signal is an example of a simple feature.
- Automated feature extraction employs specialised algorithms or deep neural networks to autonomously extract features from signals or images without the need for human intervention. This technique can be very useful when you want to develop machine learning algorithms rapidly from unprocessed data. Wavelet dispersal is an example of the automated extraction of features.

With the rise of deep learning, feature extraction has been supplanted by the initial layers of deep neural networks, but primarily for image data. For signal and time-series applications, feature extraction remains the primary obstacle that necessitates extensive knowledge prior to the development of effective predictive models.

## Notes

When you have a large data set and need to reduce the number of resources without losing essential or relevant information, the technique of extracting the features is beneficial. Feature extraction aids in the reduction of redundant data in a data set. In the end, data reduction helps to construct the model with less machine effort and increases the pace of machine learning's learning and generalisation stages.

### Applications of Feature Extraction

- **Bag of Words** - Bag-of-Words is the most popular natural language processing technique. They extract the words or characteristics from a sentence, document, website, etc. and then classify them according to their frequency of occurrence. Therefore, feature extraction is one of the most crucial components of this entire procedure.
- **Image Processing** – Image processing is one of the most innovative and intriguing fields. In this domain, you will essentially begin to experiment with your images in order to comprehend them. To process a digital image or video, we employ many techniques, including feature extraction and algorithms, to detect features such as shapes, boundaries and motion.
- **Auto-encoders** - The primary function of auto-encoders is unsupervised data coding that is efficient. This is an example of unsupervised learning. Therefore, the Feature Extraction Procedure is applicable to identify the important features of the data to be coded by learning from the original data set's coding in order to generate new ones.

## Topic 4.2 Singular Value Reduction

### Introduction

The term “Singular Value Decomposition” (SVD) refers to one of the many methods that may be used in order to cut down on the “dimensionality” (also known as the number of columns) of a data collection. Why would we want to cut down on the amount of dimensions that we have? When it comes to predictive analytics, having more columns typically implies spending more time on the modelling and scoring processes. If certain columns do not have any predictive value, this will result in time being spent, or even worse, these columns will introduce noise to the model, which will lower the quality of the model or its predicted accuracy.

It is possible to achieve dimensionality reduction by merely dropping columns, such as those that may appear to be collinear with other columns or those that are determined by an attribute importance ranking technique to not be particularly predictive of the target. Dimensionality reduction can be accomplished in this manner. But, another way to do this is to create new columns that are derived from the original columns by linear combination. In either scenario, the altered data set that was produced may be fed into machine learning algorithms, which will, in turn, provide more accurate models, quicker model building times and faster scoring times.

Dimensionality reduction is one of the possible applications for SVD; nevertheless, it is more commonly employed in digital signal processing for noise reduction, picture compression and a variety of other purposes.

An SVD is a method that factors a  $m \times n$  matrix,  $M$ , containing real or complex values into three component matrices, where the factorization takes the form  $USV^*$ . SVD may factor matrices that include either real or complex values.  $U$  is a matrix of the form  $m \times p$ .  $S$  is a  $p \times p$  diagonal matrix.  $V$  is a matrix with the dimensions  $n \times p$  and  $V^*$  is either the transpose of  $V$ , which is a matrix with the dimensions  $p \times n$ , or the conjugate transpose if  $M$  includes complex values. The value  $p$  is what is referred to as the rank. The entries that are located diagonally within  $S$  are what are known as the singular values of  $M$ . It is common practise to refer to the columns of  $U$  as the left-singular vectors of  $M$ , whereas the columns of  $V$  are commonly referred to as the right-singular vectors of  $M$ .

Dimensionality reduction refers to the process of decreasing the number of variables that are used to feed information into a predictive model. When developing a predictive model, having fewer input variables can lead to a simpler model, which may have higher performance when generating predictions based on new data. Singular Value Decomposition, or SVD for short, has become one of the most widely used methods for reducing the dimensionality of data in the field of machine learning. This method originates from the study of linear algebra and is a data preparation method that can be used to construct a projection of a sparse dataset before fitting a model. This method may be used to create a projection of a sparse dataset.

#### 4.2.1 Need for Dimensionality Reduction

The term “dimensionality reduction technique” may be described as “a strategy of transforming the higher dimensions dataset into fewer dimensions dataset ensuring that it delivers equal information.” Dimensionality reduction techniques are commonly used in statistical analysis. In the field of machine learning, these methods are frequently utilised in order to generate a predictive model that is a better match while also overcoming classification and regression issues. It is widely utilised in the industries that deal with high-dimensional data, such as voice recognition, signal processing, bioinformatics and other similar sectors. In addition to that, it may be used for things like cluster analysis, noise reduction and data visualisation.

#### Dimensionality Reduction Methods and Approaches

Now that we know the extent to which dimensionality reduction contributes to machine learning, the question becomes: what is the most effective way to carry out this process? We have provided a list of the primary techniques that you may take, further subdividing them into a variety of other ways. These strategies and procedures are collectively referred to as Dimensionality Reduction Algorithms.

##### a) Feature Selection.

The process of picking the optimal and relevant characteristics from an input data collection and deleting features that are not relevant is known as feature selection.

- **Filter methods.** A useful subset of the data set is obtained through the application of this strategy.
- **Wrapper Methods.** The performance of the characteristics that are put into this model is evaluated using this technique, which makes use of the machine learning model. It is up to the performance to decide if it is preferable to

## Notes

maintain or get rid of the characteristics in order to increase the accuracy of the model. This technique provides more precise results than filtering, but at the expense of increased complexity.

- **Embedded methods.** The machine learning model's numerous training rounds are inspected by the embedded process, which also determines the relative significance of each feature.

### b) Feature Extraction.

By this procedure, the space that possesses an excessive number of dimensions is converted into a space that possesses less dimensions. This procedure is helpful for preserving all of the information while reducing the amount of resources required for the processing of the information. The following are the three methods of extraction that are utilised most frequently.

- **Linear discriminant analysis.** Dimensionality reduction in continuous data is a typical use of the LDA technique. The data are rotated and projected in the direction of having a higher variance while using LDA. The characteristics that exhibit the greatest amount of variation are referred to as the primary components.
- **Kernel PCA.** This method is a nonlinear extension of principal component analysis (PCA) that is applicable to more complex structures that cannot be represented in a linear subspace in an easy or suitable manner. It is useful for analysing data. In order to create nonlinear mappings, KPCA makes advantage of the “kernel technique.”
- **Quadratic discriminant analysis.** This method projects the data in a way that achieves the highest possible level of class separability. Examples belonging to the same category are clustered together in the projection, whilst examples belonging to separate categories are spaced further apart.

### Dimensionality Reduction Techniques

Here are some techniques machine learning professional's use.

#### • Principal Component Analysis.

Principal component analysis, also known as PCA, is a method that is used to reduce the number of dimensions that are present in large data sets. This is accomplished by compressing a large collection of variables into a smaller set that maintains the majority of the information that was present in the original set.

Since machine learning algorithms can analyse the data much more quickly and effectively with smaller sets of information because there are fewer unnecessary factors to evaluate, accuracy must necessarily suffer as a data set's variables are reduced. This is because there are fewer unnecessary factors to evaluate. Yet, the solution to the problem of reducing the number of dimensions is to simplify things at the expense of some precision. To summarise, principal component analysis strives to keep as much information as possible while simultaneously reducing the number of variables in a data collection.

#### • Backward Feature Elimination

The performance of the model is improved by the use of backward elimination,

which begins with all of the attributes and gradually eliminates those that are of less importance. We will continue doing this until the results of removing features show no sign of improvement.

1. While first developing the model, each variable should be employed.
2. Eliminate the variable that contributes the least amount of value (for instance, based on the lowest loss in model accuracy) and then proceed with the process until a predetermined list of requirements is satisfied.

- **Forward Feature Selection**

The forward selection strategy is a procedure that involves repetitive steps and begins with the dataset containing no characteristics. The functionality of the model is improved with each iteration by the introduction of new features. In the event that performance is improved, the functionality will not be affected. Deleted characteristics include those that do not contribute to improved results. The process is repeated until there is no further improvement to be made to the model.

- **Missing Value Ratio**

Imagine the possibility of getting a dataset. Where do we even begin? Before attempting to construct a model, it is only natural that you would first study the data. Throughout your exploration of the data, you notice that some of the numbers in your dataset are absent. What should I do now? Before attempting to impute these missing values or fully eliminating the variables that have missing values, you will first investigate the root cause of the problem and find out why the values in question are absent.

What if there are too much missing data; for the sake of this discussion, let's pretend there are more than fifty percent. Should the variable be removed, or should the values that are now absent be assumed? We should not keep using the variable because it won't store too much information on its own. Nevertheless, this is not a certainty at all. We may decide to set a threshold value and if the proportion of missing data for any variable reaches that amount, we will be required to remove the variable from consideration.

- **Low Variance Filter**

The Low Variance Filter is a strategy that, like the Missing Value Ratio method, operates using a threshold. In this particular instance, though, we are checking the data columns. The approach computes the variance that is associated with each variable. Any data columns that have variances that are lower than the threshold are removed from the dataset since attributes with low variance do not have an impact on the variable of interest.

- **High Correlation Filter**

This approach is applicable to two variables that convey the same information, which may result in a reduction in the quality of the model. In this approach, we first locate the variables that have a strong correlation and then we select one of them with the help of the Variance Inflation Factor (VIF). You have the option to get rid of variables having a greater value than five ( $VIF > 5$ ).

## Notes

- **Decision Trees**

The decision tree method is a common supervised learning technique that divides data into groups that are similar depending on the variables that are supplied. This method is effective in resolving issues such as the identification of significant variables, data outliers and missing values.

- **Random Forest**

This approach is comparable to that of the decision tree strategy. On the other hand, in this circumstance, we create a huge number of trees—hence the term “forest”—against the objective variable. The next step is to locate feature subsets with the assistance of the use data provided by each attribute.

- **Factor Analysis.**

Let us suppose there are two factors at play here: one's level of education and their income. It is possible that there is a substantial connection between these two aspects, given that people with higher degrees of education typically have considerably higher salaries. The Factor Analysis method organises variables into categories according to the correlations between them; as a result, all variables that belong to the same group will have a high correlation with one another but only a moderate connection with the factors that belong to another category (s). In this context, each and every group is considered a factor. The number of these variables is rather small in relation to the original dimensions of the data. Yet, it might be challenging to spot these components.

### 4.2.2 Understanding Singular Value Reduction: Mathematical Concept

A factorization of one matrix into three additional matrices is referred to as the singular value decomposition (SVD) of that matrix. It reveals fundamental geometrical and theoretical insights regarding linear transformations, in addition to having some fascinating algebraic features and it does so in a concise manner. Also, it possesses a number of essential applications in the field of data science.

#### Mathematics behind SVD

The SVD of  $m \times n$  matrix A is given by the formula :

$$A = U W V^T$$

where:

- U:  $m \times n$  matrix of the orthonormal eigenvectors of  $A A^T$ .
- $V^T$ : transpose of a  $n \times n$  matrix containing the orthonormal eigenvectors of  $A^T A$ .
- W: a  $n \times n$  diagonal matrix of the singular values which are the square roots of the eigenvalues of  $A^T A$ .

Singular decomposition analysis(SVD)

$$C_{m \times n} = U_{m \times r} \times \Sigma_{r \times r} \times V_{r \times n}^T$$

## Examples

- Find the SVD for the matrix  $A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$
- To calculate the SVD, First, we need to compute the singular values by finding eigenvalues of  $AA^T$ .

$$A \cdot A^T = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \cdot \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix}$$

- The characteristic equation for the above matrix is:

$$\begin{aligned} W - \lambda I &= 0 \\ AA^T - \lambda I &= 0 \end{aligned}$$

$$\lambda^2 - 34\lambda + 225 = 0$$

$$= (\lambda - 25)(\lambda - 9)$$

so our singular values are:  $\sigma_1 = 5$ ;  $\sigma_2 = 3$

- Now we find the right singular vectors i.e orthonormal set of eigenvectors of  $ATA$ . The eigenvalues of  $ATA$  are 25, 9 and 0 and since  $ATA$  is symmetric we know that the eigenvectors will be orthogonal.

For  $\lambda = 25$ ,

$$A^T A - 25 \cdot I = \begin{bmatrix} -12 & 12 & 2 \\ 12 & -12 & -2 \\ 2 & -2 & -17 \end{bmatrix}$$

which can be row-reduced to:  $\begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$$

Similarly, for  $\lambda = 9$ , the eigenvector is:

$$v_2 = \begin{bmatrix} \frac{1}{\sqrt{18}} \\ \frac{-1}{\sqrt{18}} \\ \frac{4}{\sqrt{18}} \end{bmatrix}$$

For the 3rd eigenvector, we could use the property that it is perpendicular to  $v_1$  and  $v_2$  such that:

$$\begin{aligned} v_1^T v_3 &= 0 \\ v_2^T v_3 &= 0 \end{aligned}$$

Solving the above equation to generate the third eigenvector

$$v_3 = \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} a \\ -a \\ -a/2 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{-2}{3} \\ \frac{-1}{3} \end{bmatrix}$$

## Notes

## Notes

Now, we calculate U using the formula  $u_i = \frac{1}{\sigma} A v_i$  and this gives  $U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$ . Hence, our final SVD equation becomes:

$$A = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{18}} & \frac{1}{\sqrt{18}} & 0 \\ \frac{\sqrt{2}}{3} & \frac{-2}{3} & \frac{1}{3} \end{bmatrix}$$

### 4.2.3 Single Value Theorem

Basic idea

Recall from here that any matrix  $A \in R^{m \times n}$  with rank one can be written as

$$A = \sigma u v^T,$$

where  $u \in R^m, v \in R^n$  and  $\sigma > 0$ .

It turns out that a similar result holds for matrices of arbitrary rank . That is, we can express any matrix  $A \in R^{m \times n}$  with rank one as sum of rank-one matrices

$$A = \sum_{i=1}^r \sigma_i u_i v_i^T,$$

where  $u_1, \dots, u_r$  are mutually orthogonal,  $v_1, \dots, v_r$  are also mutually orthogonal and the  $\sigma_i$ 's are positive numbers called the singular values of A. In the above, r turns out to be the rank of A.

Theorem statement

The following important result applies to any matrix A and allows to understand the structure of the mapping  $x \rightarrow Ax$ .

### Theorem: Singular Value Decomposition (SVD)

An arbitrary matrix  $A \in R^{m \times n}$  admits a decomposition of the form

$$A = \sum_{i=1}^r \sigma_i u_i v_i^T = USV^T, S := \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix},$$

where  $U \in R^{m \times m}$ ,  $V \in R^{n \times n}$ , are both orthogonal matrices and the matrix S is diagonal:  $S = \text{diag}(\sigma_1, \dots, \sigma_r)$ ,

where the positive numbers  $\sigma_1 \geq \dots \geq \sigma_r > 0$  are unique and are called the singular values of A. The number  $r \leq \min(m, n)$  is equal to the rank of A and the triplet  $(U, S, V)$  is called a *singular value decomposition* (SVD) of A. The first r columns of U:  $u_i, i = 1, \dots, r$  (resp. V:  $v_i, i = 1, \dots, r$ ) are called left (resp. right) singular vectors of A and satisfy

$$Av_i = \sigma_i u_i, u_i^T A = \sigma_i v_i, i = 1, \dots, r.$$

The spectral theorem for symmetric matrices is an essential part of the demonstration that proves the theory. Note that in the theorem, the zeros appearing alongside S are really blocks of zeros. They may be empty, for example if  $r = n$ , then there are no zeros to the right of S.

## 4.3 Principal Component Analysis

Notes

### Introduction

Principal component analysis (PCA) is a popular method for analysing large datasets that contain a high number of dimensions or features for each observation. This technique improves the interpretability of data while maintaining the maximum amount of information and makes it possible to visualise multidimensional data. In its most basic form, principal component analysis (PCA) is a statistical method for decreasing the number of dimensions included within a dataset. This is performed by linearly converting the data into a new coordinate system, in which the variance in the data may be expressed using fewer dimensions than were present in the original data. This allows for a greater degree of simplification. Several studies depict the data in two dimensions using the first two principal components in order to visually detect clusters of closely linked data points. This is done using the first two principal components. Applications of principal component analysis may be found in a wide variety of domains, including population genetics, investigations of the microbiome and atmospheric research.

### 4.3.1 What is PCA?

Principal Component Analysis is a type of unsupervised learning method that is utilised in the field of machine learning for the purpose of dimensionality reduction. Using orthogonal transformation, this statistical method transforms the observations of correlated characteristics into a collection of linearly uncorrelated data. This is done by converting the correlated features into orthogonal coordinates. These newly remodelled characteristics have been given the name Main Components. It is one of the most common tools that is utilised in the process of exploratory data analysis and predictive modelling. It is a method for extracting robust patterns from a given dataset by taking steps to minimise the amount of variation in the data. In most cases, principal component analysis will look for a lower-dimensional surface onto which to project higher-dimensional data.

PCA works by taking into account the variance of each characteristic. This is done because a high attribute demonstrates a good split between the classes, which in turn minimises the dimensionality of the problem. Image processing, movie recommendation systems and optimising power distribution across a variety of communication channels are some real-world uses of principal component analysis (PCA). Because it is a method for extracting features, it takes into account the most significant factors while disregarding the less significant ones.

The PCA method is founded on a number of mathematical ideas, including the following:

- Variance and Covariance
- Eigenvalues and Eigen factors

The following is a list of some popular terminology used in the PCA algorithm:

**Dimensionality:** It refers to the total number of characteristics or variables that are contained inside a certain dataset. The number of columns that are included in the dataset is a much simpler indicator of this.

## Notes

- **Correlation:** The term “correlation” refers to the degree to which two different variables are connected to one another. For example, if one variable is updated, it will cause the other variable to likewise change. The correlation value might be anything from minus one to plus one. In this case, we get a value of -1 if the variables in question are inversely proportional to one another and we get a value of +1 if the variables in question are directly proportional to one another.
- **Orthogonal:** This hypothesis states that the variables do not have any kind of relationship with one another and as a result, there is no correlation between the two sets of variables.
- **Eigenvectors:** In the case when there is a square matrix  $M$  and a vector that is not zero is provided. If a scalar multiple of  $v$ ,  $Av$ , exists, then  $v$  will be an eigenvector in that case.
- **Covariance Matrix:** The term “covariance matrix” refers to a matrix that contains information on the covariance that exists between two variables.

### Principal Components in PCA

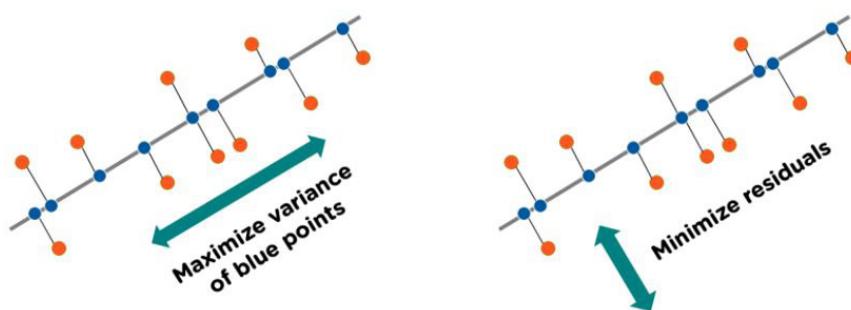
As was just said, the Principal Components are either the newly modified features or the results of the principal component analysis. The number of these PCs is either less than the total number of the original characteristics that were included in the dataset or it is the same. The following is a list of some of the characteristics of these primary components:

- The linear combination of the initial features needs to be the primary component.
- These components are considered orthogonal, which means that there is no correlation between any two of the variables being considered.
- Since the significance of each component lessens as the number of components increases from one to  $n$ , the relevance of component number one is greatest and the significance of component number  $n$  is lowest.

### Applications of Principal Component Analysis

- The principal function of principal component analysis (PCA) in artificial intelligence (AI) applications like computer vision and image compression, amongst others, is to perform dimensionality reduction.
- PCA can also be utilised for the discovery of hidden patterns when the data in question has a high number of dimensions. PCA is utilised in a variety of industries, including finance, data mining, psychology and others.

### How Does Principal Component Analysis Work?



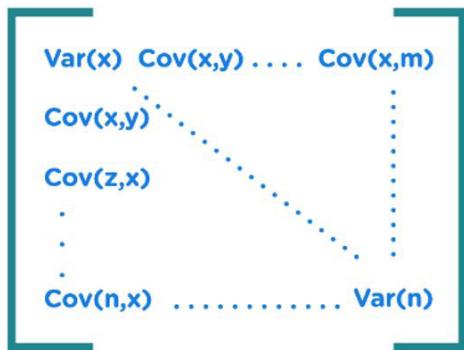
## 1. Normalize the Data

Before beginning the PCA analysis, standardise the data. This will guarantee that each characteristic has a mean value of zero and a variance value of one.

$$Z = \frac{x - \mu}{\sigma}$$

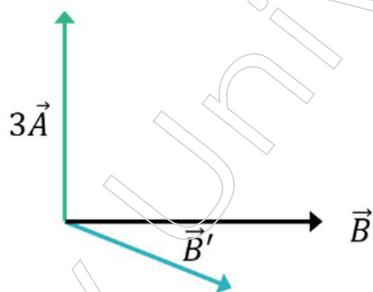
## 2. Build the Covariance Matrix

Build a square matrix in order to represent the correlation between two or more characteristics in a dataset that contains several dimensions.



## 3. Find the Eigenvectors and Eigenvalues

Do the calculations necessary to determine the eigenvalues and eigenvectors/unit vectors. Scalars are used to multiply the eigenvector of the covariance matrix and eigenvalues are one type of scalar.



## 4. Sort the Eigenvectors in Highest to Lowest Order and Select the Number of Principal Components.

Let us carry out a hands-on demonstration on principal component analysis using Python now that you have a better understanding of how PCA works in machine learning.

### Advantages of PCA

PCA offers a multitude of advantages when it comes to the analysis of data, including the following:

- Dimensionality reduction:** Reduction in dimensionality One of the key advantages of principal component analysis is that it helps decrease the dimensionality of the data by identifying the characteristics or components that are most important. When the initial data comprises a large number of variables and is thus difficult to view or interpret, this might be of great assistance.

## Notes

2. **Feature Extraction:** Principal component analysis (PCA) may also be used to generate additional features or elements from the original data. These new features or elements might be more insightful or intelligible than the original features. This is especially effective in situations in which the initial characteristics are associated with one another or noisy.
3. **Data visualization:** Principal component analysis (PCA) may be used to show high-dimensional data in two or three dimensions by projecting the data onto the first few principal components. This can be helpful in discovering data patterns or clusters that would not have been obvious in the initial high-dimensional space that was being used.
4. **Noise Reduction:** PCA may also be utilised to minimise the effects of noise or measurement mistakes in the data by detecting the underlying signal or pattern in the data. This is accomplished through the process of "noise reduction."
5. **Multicollinearity:** If two or more variables are highly associated with one another, then the data include multicollinearity, which PCA is able to account for and handle. By determining which characteristics or components are the most important, principal component analysis (PCA) can help mitigate the effects of multicollinearity on an analysis.

### Disadvantages of PCA

1. **Interpretability:** Although though principle component analysis (PCA) is an efficient method for decreasing the dimensionality of data and identifying patterns, the principal components that are produced as a consequence of the analysis are not always easy to comprehend or define in terms of the qualities they were based on.
2. **Information loss:** Principal component analysis (PCA) requires selecting a subset of the most important characteristics or components in order to decrease the dimensionality of the data. Even while this can be beneficial for organising the data and reducing noise, there is a possibility that information could be lost if some essential traits were omitted from the components that were selected.
3. **Outliers:** Since PCA is vulnerable to irregularities in the data, the primary components that are generated as a consequence may be considerably influenced. Outliers have the potential to skew the covariance matrix, which can make it more difficult to determine which traits are the most important.
4. **Scaling:** The principal component analysis (PCA) operates under the premise that the data is scaled and centralised, which might be a limitation in some contexts. If the data are not scaled appropriately, the resultant main components may not accurately portray the underlying patterns in the data. If the data are not scaled properly.
5. **Computing complexity:** In the case of large datasets, the computation of the eigenvectors and eigenvalues of the covariance matrix may be expensive. It's possible that this will limit PCA's capacity to scale, rendering it ineffective for some applications.

### 4.3.2 Algorithm for PCA

Principal Component Analysis is a technique that helps you determine which aspects of your project are most similar to one another and simplifies the process of

result analysis. Imagine for a moment that you are working on a project that has a substantial number of different variables and dimensions. There will be some of these factors that are more important than others. While others might not be major important variables, some might be. Hence, the Principal Component Method of component analysis provides you with a calculative means of removing a few additional variables that are not as relevant, thereby preserving the openness of all of the information. Is there any way around this? The answer is yes; this is doable. Because of this, dimensionality reduction is another name for the technique known as principal component analysis. You may quickly investigate and visualise the methods by reducing the amount of data and dimensions involved and you will not have to waste any of your precious time doing so. As a result, principal component analysis (PCA) statistics is the science of assessing all of the dimensions and decreasing them as much as feasible while maintaining the precise information. Where Can You Find Applications of Principal Component Analysis in Python and Machine Learning? The following list provides access to some of the PCA's application options.

- You can monitor multi-dimensional data (can visualise in 2D or 3D dimensions) over any platform by using the Principal Component Method of factor analysis.
- PCA techniques help with data cleaning and data pre-processing techniques.
- PCA techniques aid in data cleaning and data pre-processing techniques.
- PCA is able to assist in the compression of data and the transmission of that data by utilising efficient PCA analysis methods. All of these different methods for processing information don't compromise the quality in any way.
- This statistic is the science of assessing different dimensions and it has the potential to be employed in a variety of contexts, such as face recognition, picture identification, pattern recognition and a great deal more besides.
- The principal component analysis (PCA) approach used in machine learning helps to simplify complicated business processes.
- Due to the fact that Principal Component Analysis reduces the variance of dimensions that is more relevant, it is simple to denoise the information and fully omit the noise as well as any external influences.

#### Steps for PCA algorithm

1. **Getting the dataset:** First things first, we need to take the dataset that was provided to us and split it into two parts: X and Y. X will serve as our training set, while Y will serve as our validation set.
2. **Representing data into a structure:** The next step is to create a structure that represents our dataset. In this manner, the two-dimensional matrix of the independent variable X will be represented by us. In this table, each row represents a different data item and each column represents a different feature. The dimensions of the dataset are indicated by the number of columns.
3. **Standardizing the data:** The standardisation of our dataset will take place in this stage. For example, in a certain column, the elements that have a higher variation are considered to be more essential than the features that have a smaller variance. If the significance of characteristics is not reliant on the degree to which those features

## Notes

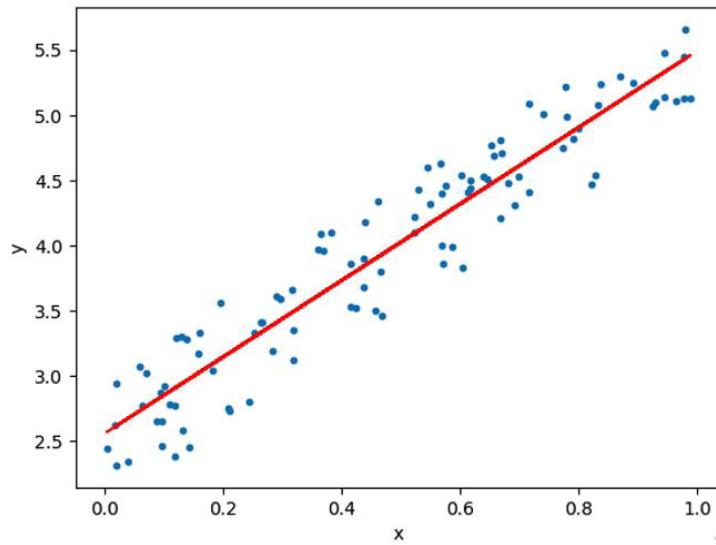
vary, then we shall divide each data item in a column by the column's standard deviation. From this point, we shall refer to the matrix as Z.

4. **Calculating the Covariance of Z:** In order to compute the covariance of Z, we will begin by transposing the matrix Z that we have just created. Following the transposition, we shall carry out the multiplication by Z. The covariance matrix of Z is going to be the matrix that is output.
5. **Calculating the Eigen Values and Eigen Vectors:** Now that we have the resulting covariance matrix Z, we need to determine its eigenvalues and eigenvectors. The directions of the axes that contain the most information are referred to as eigenvectors or the covariance matrix. Moreover, the eigenvalues are referred to as the coefficients of these eigenvectors.
6. **Sorting the Eigen Vectors:** During this stage of the process, we will collect all of the eigenvalues and arrange them in descending order, which means from the greatest to the least significant. And at the same time, arrange the eigenvectors in the appropriate order in the matrix P of eigenvalues.  $P^*$  will be the name given to the final matrix that was created.
7. **Calculating the new features Or Principal Components:** In this section, we will compute the newly added characteristics. In order to do this, we are going to multiply the  $P^*$  matrix by the Z. Each observation in the resulting matrix  $Z^*$  is the linear combination of the characteristics that were present in the original data. Independent of one another, the columns of the  $Z^*$  matrix can be arranged in any order.
8. **Remove less or unimportant features from the new dataset:** The new feature set has been implemented, therefore from this point on, we will select what to keep and what to get rid of. That implies that in the new dataset, we will only maintain the characteristics that are relevant or important and we will eliminate the features that are not relevant or essential.

### 4.3.3 Application and Role of PCA in Dimensionality Reduction

One of the most common types of linear dimension reduction techniques is known as principal component analysis (PCA). It is a projection-based approach that changes the data by projecting it onto a set of orthogonal (perpendicular) axes. This method has been around for quite some time.

"The principal component analysis works on the assumption that when the data in a higher-dimensional space are translated to data in a lower-dimensional space, the variance or spread of the data in the lower-dimensional space should be as great as possible." The data in the picture below have the greatest amount of variation along the red line, which represents two-dimensional space.

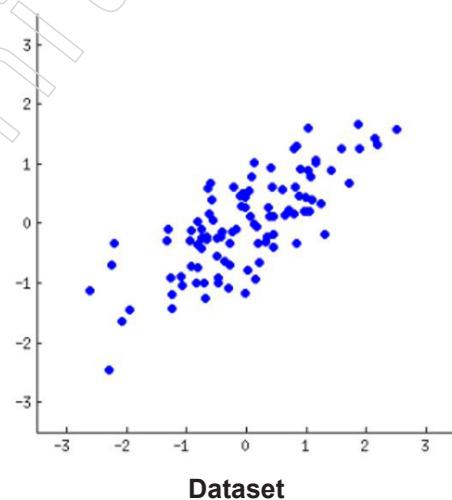
**Notes**

### Intuition

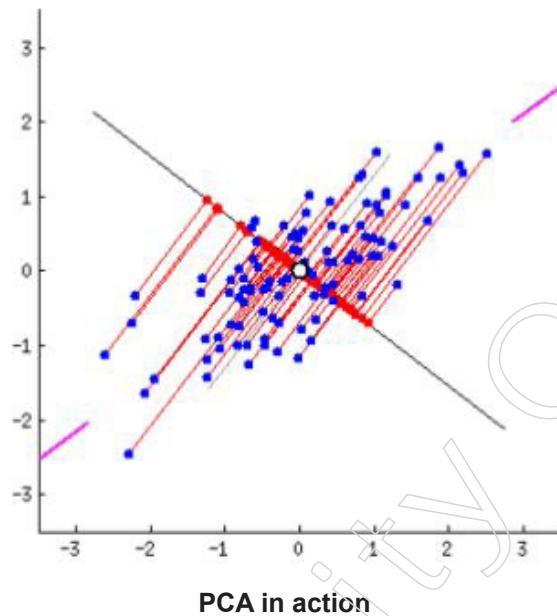
Let us develop an intuitive understanding of PCA. Let us say that you want to differentiate between various kinds of food depending on the amount of nutrients that they contain. In order to separate the various food products, which variable would be the best choice? If you select a variable that differs substantially from one food item to another, then you will be able to appropriately remove them from one another. If the chosen variable is almost present in the same quantity in each food item, your task will be made much more difficult. What if the data doesn't contain a variable that correctly separates the different types of food? It is possible for us to generate a new variable by using a linear combination of the variables that are already there, such as:

$$\text{New\_Var} = 4 * \text{Var1} - 4 * \text{Var2} + 5 * \text{Var3}.$$

PCA searches for the linear combinations of the original variables that provide the greatest amount of variation or spread along the new variable and it achieves this by finding the optimal linear combinations. Let us say we need to convert a representation of data points in two dimensions into a representation in one dimension. As a result, we will search for a straight line and attempt to plot data points along it. (There is just one dimension in a straight line). There are many different options available to choose a straight line.



## Notes



Consider that the magenta line will serve as the new dimension for us. If you can see the red lines (which link the projection of blue dots on a magenta line), this indicates that the projection error is equal to the perpendicular distance that each data point is from the straight line. The overall projection error will be equal to the sum of the errors associated with all of the data points.

Those first data points, which were blue, will serve as the basis for our new data points, which will be projections of those points. By projecting the data points onto a one-dimensional space, in the shape of a straight line, we have, as can be seen, reduced the number of dimensions that our data points occupy from two to one. The name for this crimson line that runs straight through the middle is the primary axis. While we are just projecting into a single dimension, we only have a single primary axis to work with. In order to identify the subsequent primary axis based on the residual variance, we use the same process. The next principal axis, in addition to being the direction in which there is the greatest amount of variation, must also be orthogonal, or perpendicular or uncorrelated to each other, to the other principal axes.

When all of the primary axes have been determined, the dataset will next be projected onto those axes. Principal components are the columns that remain in the dataset after it has been projected or modified.

The principal components are essentially the linear combinations of the original variables; the weights vector in this combination is actually the eigenvector that was found, which in turn satisfies the principle of least squares. The principal components can be found by using the principal component analysis method.

Because linear algebra exists, we do not have to worry too much about the principal component analysis (PCA). Eigenvalue Decomposition and Singular Value Decomposition (SVD), both of which originate in linear algebra, are the two primary processes that are employed in principal component analysis (PCA) to minimise the number of dimensions.

## Eigenvalue Decomposition

**Matrix decomposition** is a procedure that breaks down a matrix into its component elements in order to ease a variety of tasks that would otherwise be extremely difficult. Decomposing a square matrix ( $n$  by  $n$ ) into a collection of eigenvectors and eigenvalues is the process known as eigenvalue decomposition and it is the matrix decomposition approach that is used the most frequently.

**Eigenvectors** are considered to be unit vectors, which indicates that the length or magnitude of an eigenvector is always equal to 1.0. It is common practise to refer to them as right vectors, which euphemistically stands for column vectors (as opposed to a row vector or a left vector).

The length or size of an eigenvector is determined by the coefficients that are applied to it, which are known as eigenvalues. For instance, when scaling it, a negative eigenvalue might cause the direction of the eigenvector to change in the opposite way.

According to the rules of mathematics, a vector is considered to be an eigenvector of any  $n \times n$  square matrix  $A$  if and only if it fulfils the following equation:

$$A \cdot v = \lambda \cdot v$$

This equation is known as the eigenvalue equation, where  $A$  is the  $n \times n$  parent square matrix that we are deconstructing,  $v$  is the matrix's eigenvector and  $\lambda$  represents the eigenvalue scalar.

In simpler words, the linear transformation of a vector  $v$  by  $A$  has the same effect of scaling the vector by factor  $\lambda$ . Note that for  $m \times n$  non-square matrix  $A$  with  $m \neq n$ ,  $A \cdot v$  an  $m$ -D vector but  $\lambda \cdot v$  is an  $n$ -D vector, i.e., no eigenvalues and eigenvectors are defined. If you wanna dive deeper into mathematics check this out.

|                                                  |                                                  |                                                  |                                                  |
|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| Original<br>Matrix                               | Eigenvectors<br>Matrix                           | Eigenvalues<br>Matrix                            |                                                  |
| $\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$ | $\begin{bmatrix} -1 & -1 \\ 2 & 1 \end{bmatrix}$ | $\begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}$ | $\begin{bmatrix} 1 & 1 \\ -2 & -1 \end{bmatrix}$ |
|                                                  |                                                  |                                                  | Inverse of<br>Eigenvectors<br>Matrix             |

## Eigenvalue Decomposition

The original matrix can be reconstructed by multiplying all of its constituent matrices together, or by combining the transformations that are represented by the matrices themselves. The matrix is not compressed because of a decomposition process; rather, the matrix is broken down into its component components to make it simpler to conduct specific operations on the matrix. Eigen decomposition, much like other methods of matrix decomposition, can be utilised as a component to simplify the computation of other, more difficult matrix operations.

## Notes

### 4.3.4 Example for Finding PCA in Dataset

Principal Component Analysis is essentially a statistical process that is used to transform a set of observations of potentially correlated variables into a set of values of linearly uncorrelated variables. This is accomplished via the use of a principal component.

Each of the principle components was selected in such a manner that it would be able to characterise the majority of the variation that was still accessible and all of these principal components are orthogonal to one another. Among all principal components first principal component has a highest variance.

#### Uses of PCA:

- The purpose of this method is to determine which of the variables in the data are related to one another.
- It allows for the interpretation and visualisation of data.
- Since there are fewer factors to take into account, subsequent analysis will be less complicated.
- It is frequently utilised to depict the genetic distance between populations as well as their relatedness to one another.

These operations are carried out, fundamentally, on a square symmetric matrix. It is possible for this matrix to be a covariance matrix, a correlation matrix, or a pure sums of squares and cross-products matrix. In cases where the individual variances differ greatly from one another, a correlation matrix is utilised.

#### Objectives of PCA:

- At its core, it is an independent process that, among other things, narrows the attribute space by reducing the number of variables and factors to which it is subject to consideration.
- Principal component analysis is essentially a technique of dimension reduction; however, there is no assurance that the dimensions can be interpreted.
- The primary objective of this principle component analysis (PCA) is to determine which of the original variables have the highest correlation with the principal amount in order to choose a subset of variables from a broader collection of variables.

**Principal Axis Method:** PCA looks, at its core, for a linear combination of variables in order to get the greatest possible amount of variance from those variables. As soon as this procedure is finished, it gets rid of it and starts looking for another linear combination that explains the highest proportion of residual variance, which ultimately leads to orthogonal components. We do an analysis of the total variance using this approach.

**Eigenvector:** It is a vector that is not zero and maintains its parallelism after the matrix is multiplied. Let us say that  $Mx$  and  $x$  are parallel. In this case, we will assume that  $x$  is an eigenvector of size  $r$  of matrix  $M$ , which has dimension  $r \times r$ . Consequently, in order to obtain the eigenvector and the eigenvalues, we need to solve the equation  $Mx = Ax$ , where both  $x$  and  $A$  are unknown.

We may state that Principal Components indicate both the Common and Unique Variance of the Variable under the Eigen-Vectors heading. In its most basic form, it is a strategy that focuses on variance and aims to duplicate the overall variance as well as the correlation with all components. The principal components are essentially the linear combinations of the initial variables, with each variable's contribution to the total variance in a specific orthogonal dimension serving as the weighting factor for the main component.

**Eigen Values:** The term “characteristic roots” refers to this concept in its most basic form. In essence, it assesses the amount of variation across all variables that can be attributed to that one factor. The ratio of eigenvalues may be thought of as the ratio of the relevance of the factors in explaining the variables to the importance of the variables themselves. When the value of the component is low, it makes a less contribution to the overall explanation of the variables. To put it another way, it determines how much of the whole provided database's variation may be attributed to the element being measured. The eigenvalue of the factor may be determined by taking the total of its squared factor loadings for each of the variables into account.

Now, let us figure out how to use Python to understand principal component analysis.

#### Step 1: Importing the library files

- Python
- ```
# importing required libraries  
  
import numpy as np  
  
import matplotlib.pyplot as plt  
  
import pandas as pd
```

Step 2: The second step is importing the data set.

Import the dataset and then break it up into its X and y components so that it can be analysed.

- Python
- ```
importing or loading the dataset

dataset = pd.read_csv('wine.csv')

distributing the dataset into two components X and Y

X = dataset.iloc[:, 0:13].values

y = dataset.iloc[:, 13].values
```

#### Step 3: The third step involves dividing the dataset into the Training set and the Test set.

- Python
- ```
# Splitting the X and Y into the  
  
# Training set and Testing set
```

Notes

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

Step 4: Feature Scaling

Doing the steps involved in the preprocessing of the training and testing set, such as fitting the Standard scale.

- Python

```
# performing preprocessing part
```

```
from sklearn.preprocessing import StandardScaler
```

```
sc = StandardScaler()
```

```
X_train = sc.fit_transform(X_train)
```

```
X_test = sc.transform(X_test)
```

Step 5: Applying PCA function

The PCA function was applied to both the training set and the testing set in order to conduct analysis.

- Python

```
# Applying PCA function on training
```

```
# and testing set of X component
```

```
from sklearn.decomposition import PCA
```

```
pca = PCA(n_components = 2)
```

```
X_train = pca.fit_transform(X_train)
```

```
X_test = pca.transform(X_test)
```

```
explained_variance = pca.explained_variance_ratio_
```

Step 6: Fitting Logistic Regression To the training set

- Python

```
# Fitting Logistic Regression To the training set
```

```
from sklearn.linear_model import LogisticRegression
```

```
classifier = LogisticRegression(random_state = 0)
```

```
classifier.fit(X_train, y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
penalty='l2', random_state=0, solver='liblinear', tol=0.0001,
verbose=0, warm_start=False)
```

Step 7: Speculating how the test set will turn out.

- Python

Notes

```
# Predicting the test set result using  
# predict function under LogisticRegression  
y_pred = classifier.predict(X_test)
```

Step 8: Creating the confusion matrix is the eighth step.

- Python

```
# making confusion matrix between  
# test set of Y and predicted value.  
from sklearn.metrics import confusion_matrix  
cm = confusion_matrix(y_test, y_pred)
```

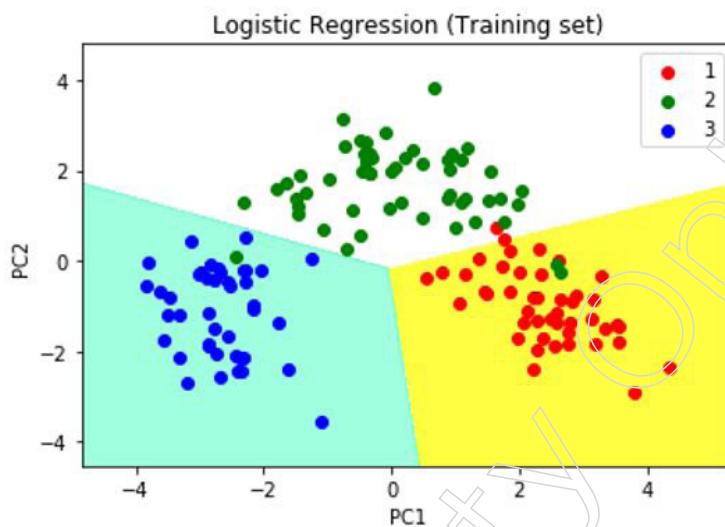
Step 9: Predicting how the training set will turn out.

- Python

```
# Predicting the training set  
# result through scatter plot  
from matplotlib.colours import ListedColourmap  
X_set, y_set = X_train, y_train  
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1,  
                               stop = X_set[:, 0].max() + 1, step = 0.01),  
                               np.arange(start = X_set[:, 1].min() - 1,  
                               stop = X_set[:, 1].max() + 1, step = 0.01))  
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(),  
                                                 X2.ravel()]).T).reshape(X1.shape), alpha = 0.75,  
cmap = ListedColourmap(('yellow', 'white', 'aquamarine')))  
plt.xlim(X1.min(), X1.max())  
plt.ylim(X2.min(), X2.max())  
for i, j in enumerate(np.unique(y_set)):  
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],  
                c = ListedColourmap(('red', 'green', 'blue'))(i), label = j)  
plt.title('Logistic Regression (Training set)')  
plt.xlabel('PC1') # for xlabel  
plt.ylabel('PC2') # for ylabel  
plt.legend() # to show legend  
# show scatter plot
```

Notes

```
plt.show()
```

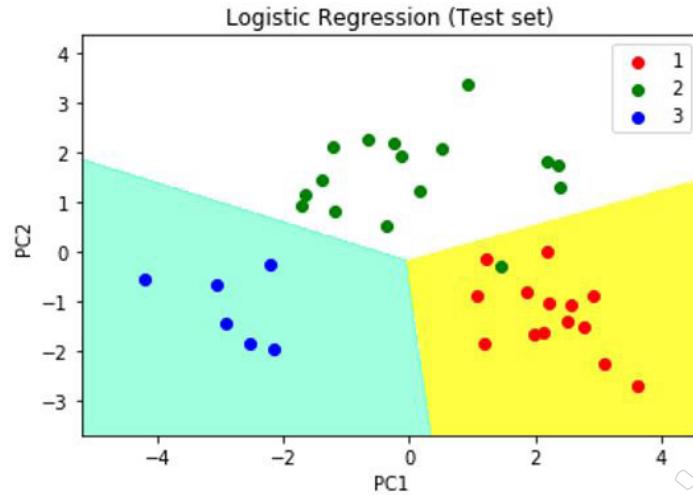


Step 10: Visualizing the Test set results.

- Python

```
# Visualising the Test set results through scatter plot
from matplotlib.colours import ListedColourmap
X_set, y_set = X_test, y_test
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1,
                                 stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1,
                               stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(),
                                                 X2.ravel()]).T).reshape(X1.shape), alpha = 0.75,
             cmap = ListedColourmap(('yellow', 'white', 'aquamarine')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColourmap(('red', 'green', 'blue'))(i), label = j)
# title for scatter plot
plt.title('Logistic Regression (Test set)')
plt.xlabel('PC1') # for xlabel
plt.ylabel('PC2') # for ylabel
plt.legend()
```

```
# show scatter plot
plt.show()
```



In the new principal component space, we can visualise what the data looks like:

- Python3

```
# plot the first two principal components with labels
y = df.iloc[:, -1].values
colours = ["r", "g"]
labels = ["Class 1", "Class 2"]
for i, colour, label in zip(np.unique(y), colours, labels):
    plt.scatter(X_pca[y == i, 0], X_pca[y == i, 1], colour=colour, label=label)
plt.xlabel("Principal Component 1")
plt.ylabel("Principal Component 2")
plt.legend()
plt.show()
```

7

This is a very simple example of how to use Python to do PCA. This code will make a scatter plot of the first two principal components and the ratio of their explained variance. By choosing the right number of principal components, we can cut down on the number of dimensions in the dataset and better understand it.

Summary

- Dimensionality refers to the number of input features, variables, or columns present in a given dataset and dimensionality reduction refers to the process of reducing these features.
- A dataset contains a large number of input features in various circumstances, which complicates the task of predictive modelling.

Notes

Notes

- Dimensionality reduction technique can be defined as “the process of transforming a dataset with greater dimensions into a dataset with fewer dimensions while preserving the same information.”
- Predictive modelling is a probabilistic method for forecasting outcomes based on a set of predictors. These predictors are essentially characteristics that play a role in determining the model’s ultimate outcome.
- The forward selection strategy is a procedure that involves repetitive steps and begins with the dataset containing no characteristics.
- High Correlation Filter is an approach applicable to two variables that convey the same information, which may result in a reduction in the quality of the model. In

Glossary

- **Bag of Words** - Bag-of-Words is the most popular natural language processing technique. They extract the words or characteristics from a sentence, document, website, etc.
- **Image Processing** – Image processing is one of the most innovative and intriguing fields. In this domain, you will essentially begin to experiment with your images in order to comprehend them.
- **Auto-encoders** - The primary function of auto-encoders is unsupervised data coding that is efficient. This is an example of unsupervised learning.
- **Dimensionally Reduction:** The term “dimensionality reduction technique” may be described as “a strategy of transforming the higher dimensions dataset into fewer dimensions dataset ensuring that it delivers equal information.”
- **Kernel PCA.** This method is a nonlinear extension of principal component analysis (PCA) that is applicable to more complex structures that cannot be represented in a linear subspace in an easy or suitable manner.
- **Quadratic discriminant analysis.** This method projects the data in a way that achieves the highest possible level of class separability.

Check Your Understanding

1. What is the full form of PCA?
 - a) Principal Component Analysis
 - b) Partial Component Analysis
 - c) Pattern Component Analysis
 - d) Pearson Component Analysis
2. What is the full form of SVD?
 - a) Singular Value Decomposition
 - b) Simple Value Decomposition
 - c) Sample Value Decomposition
 - d) Symmetric Value Decomposition

3. What is the full form of LDA?
 - a) Linear Discriminant Analysis
 - b) Level Discriminant Analysis
 - c) Level Division Analysis
 - d) Linear Division Analysis
4. What is the full form of GLM?
 - a) Global Linear Model
 - b) Generalised Linear Multicollinearity
 - c) Global Level Model
 - d) Generalized Linear Model
5. What is the full form of GANs?
 - a) Global Adversarial Networks
 - b) Generative Adversarial Networks
 - c) Generalised Adversarial Networks
 - d) Generalised Application Networks
6. What is the the full form of LIME?
 - a) Less-interpretable-model-agnostic explanations
 - b) Local-interpretable-model-agnostic explanations
 - c) Local-interpretable-model-analysis explanations
 - d) Local-interpretable-model-agnostic examples
7. _____ are a prevalent algorithm for supervised learning that divides data into homogeneous categories based on input variables.
 - a) Principal Component Analysis
 - b) Machine Learning Algorithm
 - c) Decision trees
 - d) Eigenvalue Decomposition
8. _____ techniques can be applied to labelled data in order to identify the most important features for improving the performance of supervised models such as classification and regression.
 - a) Unsupervised Methods
 - b) Supervised Techniques
 - c) Structured Techniques
 - d) Semi-structured Techniques
9. _____ methods can be applied to unlabelled data. Examples include K-Means Clustering, Principal Component Analysis and Hierarchical Clustering. Here, we must identify a subset of the initial set of variables.

Notes

- a) Unsupervised Methods
 - b) Text Mining Techniques
 - c) Filter Methods
 - d) Wrapper Methods
10. The term _____ (SVD) refers to one of the many methods that may be used in order to cut down on the “dimensionality” (also known as the number of columns) of a data collection.
- a) Principal Component Analysis
 - b) Eigenvalue Decomposition
 - c) Singular Value Decomposition
 - d) Eigenvectors
11. The term _____ may be described as “a strategy of transforming the higher dimensions dataset into fewer dimensions dataset ensuring that it delivers equal information.”
- a) Missing Value Ratio
 - b) Dimensionality Reduction Technique
 - c) Forward Feature Selection
 - d) Backward Feature Elimination
12. _____ is a nonlinear extension of principal component analysis (PCA) that is applicable to more complex structures that cannot be represented in a linear subspace in an easy or suitable manner.
- a) Kernel PCA
 - b) Linear discriminant analysis
 - c) Quadratic discriminant analysis
 - d) Embedded Technique
13. _____, also known as PCA, is a method that is used to reduce the number of dimensions that are present in large data sets.
- a) Backward Feature Elimination
 - b) Principal Component Analysis
 - c) Quadratic discriminant analysis
 - d) Missing Value Ratio
14. The term _____ refers to the degree to which two different variables are connected to one another.
- a) Correlation
 - b) Regression
 - c) Linear Regression
 - d) Multi Regression

Notes

15. The term _____ refers to a matrix that contains information on the covariance that exists between two variables.
- a) Eigenvectors
 - b) Correlation
 - c) Covariance Matrix
 - d) Orthogonal
16. _____ is a procedure that breaks down a matrix into its component elements in order to ease a variety of tasks that would otherwise be extremely difficult.
- a) Regularisation
 - b) Matrix decomposition
 - c) Random Forest
 - d) Chi-Square
17. _____ are considered to be unit vectors, which indicates that the length or magnitude of an eigenvector is always equal to 1.0.
- a) Outliners
 - b) Scalers
 - c) Eigenvectors
 - d) Complexities
18. _____ hypothesis states that the variables do not have any kind of relationship with one another and as a result, there is no correlation between the two sets of variables.
- a) Orthogonal
 - b) Eigenvectors
 - c) Correlation
 - d) Covariance Matrix
19. _____ method is a nonlinear extension of principal component analysis (PCA) that is applicable to more complex structures that cannot be represented in a linear subspace in an easy or suitable manner.
- a) Linear discriminant analysis
 - b) Quadratic discriminant analysis
 - c) Kernel PCA
 - d) Covariance Matrix
20. In _____, feature selection is approached as a search problem in which various combinations are generated, evaluated and compared to other combinations. It trains the algorithm iteratively using the subset of features.
- a) Supervised Methodology

Notes

- b) Wrapper Methodology
- c) Embedded Methodology
- d) Filter Methodology

Exercise

1. Explain different components of Dimensionality Reduction.
2. Explain dimensionality reduction methods and approaches.
3. Explain various dimensionality reduction techniques.

Learning Activities

1. Explain the concept of Predictive Modelling.
2. Explain the concept of Principal Component Analysis.

Check your Understanding – Answers

- | | |
|--------|--------|
| 1. a) | 2. a) |
| 3. a) | 4. d) |
| 5. b) | 6. b) |
| 7. c) | 8. b) |
| 9. a) | 10. c) |
| 11. b) | 12. a) |
| 13. b) | 14. a) |
| 15. c) | 16. b) |
| 17. c) | 18. a) |
| 19. c) | 20. b) |

Further Readings and Bibliography

1. <https://towardsdatascience.com/mcculloch-pitts-model-5fdf65ac5dd1>
2. <https://www.geeksforgeeks.org/activation-functions-neural-networks/>
3. <https://www.techtarget.com/searchenterpriseai/definition/backpropagation-algorithm>
4. <https://www.geeksforgeeks.org/backpropagation-in-data-mining/>
5. https://www.sas.com/en_in/insights/analytics/neural-networks.html#:~:text=Neural%20networks%20are%20computing%20systems,time%20%E2%80%93%20continuously%20learn%20and%20improve

Module -V: Text Analysis and Retrieval

Notes

Learning Objectives

At the end of this topic, you will be able to understand:

- Analyse Introduction to Text mining
- Identify definition and language for data science
- Describe collection of data-hunting, logging, scrapping
- Identify cleaning data-artifacts, data compatibility
- Analyse dealing with missing values, outliers
- Describe definition, evolution of big data and its importance
- Analyse four Vs in big data, drivers for big data
- Learn about big data analytics, big data applications
- Identify designing data architecture, R syntax
- Describe IDE for hadoop, integration with big data, integration methods
- Interpret introduction to neural network
- Analyse difference between human brain and artificial network
- Analyse perceptron model: its features, McCulloch Pits model
- Identify role of activation function, backpropagation algorithm
- Analyse neural network in data science
- Describe role of FAT in data science, ethical challenges in data science
- Analyse some real-life examples: Covid19, data breach cases

Introduction

Text mining is the process of exploring and analysing huge volumes of unstructured text data with the assistance of software that can recognise concepts, patterns, subjects, keywords and other properties included within the data. This exploration and analysis process is known as text mining. Text analytics is another name for it, however some people differentiate between the two phrases. According to one viewpoint, text analytics refers to the application that makes use of text mining techniques in order to filter through data sets.

5.1 Text Mining and Information Retrieval

Information Retrieval (IR) is a software programme that deals with the organisation, storage, retrieval and assessment of information from document repositories, particularly textual information. IR may also be described as a term that refers to the process of retrieving information. Information Retrieval is the activity of obtaining content that can typically be documented on an unstructured nature, i.e., typically text, that satisfies an information need from within large collections that are stored on computers. This material can be obtained through the use of information retrieval

Notes

systems. One scenario that falls under the category of “Information Retrieval” is when a user submits a query to a database.

5.1.1 Introduction to Text Mining

Text mining, also known as text data mining, is the process of extracting significant patterns and fresh insights from unstructured text by converting the material into a structured format. Companies are able to investigate and identify hidden links within their unstructured data when they employ advanced analytical approaches such as Naive Bayes, Support Vector Machines (SVM) and other deep learning algorithms. Inside databases, text is one of the forms of data that is used the most frequently. This information could be arranged in the following ways, depending on the database:

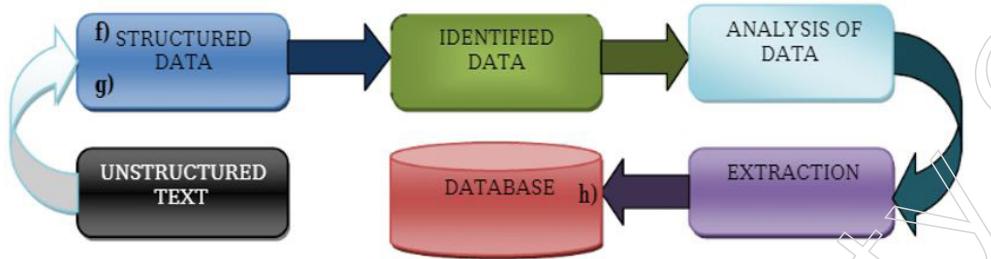
- **Structured data:** This data is standardised into a tabular structure with several rows and columns, which makes it easier to store and process for analysis and machine learning algorithms. This data is also referred to as “clean data.” Inputs like names, addresses and phone numbers are all examples of the kinds of things that can be included in structured data.
- **Unstructured data:** This kind of data does not adhere to any particular data format that has already been established. Text from various sources, such as social media or product reviews, as well as rich media formats, such as video and audio files, may be included in this section.
- **Semi-structured data:** As the name implies, this type of data is a combination of structured and unstructured data formats. The data in this type can be organised in a variety of ways. Although it is organised to some degree, it does not possess the necessary level of structure to fulfil the prerequisites of a relational database. Files written in XML, JSON and HTML are all examples of types of data that are considered semi-structured.

Text mining is an immensely helpful activity for businesses to implement due to the fact that the majority of data in the world is stored in an unstructured manner. Text mining tools and natural language processing (NLP) approaches, such as information extraction (PDF, 131 KB) (link resides outside of IBM), enable us to transform unstructured materials into a structured format, which in turn enables analysis and the development of high-quality insights. This, in turn, leads to improved decision-making inside companies, which in turn leads to improved outcomes for businesses.

The following is a rundown of the five primary steps involved in text mining:

- Collecting unstructured data from a variety of data sources, including but not limited to plain text, web pages, pdf files, emails and blogs, to mention a few of these sources.
- Conduct pre-processing and purification procedures on the data in order to identify and eliminate any abnormalities that may be present. The process of data purification enables you to retrieve and store the useful information that is buried within the data and assists in determining the origins of phrases.
- In exchange for this, you will receive a variety of text mining applications as well as text mining tools.

- Transform into structured forms all of the pertinent information that was taken from the unstructured data.
- Using the Management Information System, do an analysis of the patterns present within the data (MIS).
- Place all of the important information into a protected database so that trend analysis can be performed and the organisation's decision-making process may be improved.



Text Mining Techniques

The strategies of text mining may be comprehended by looking at the procedures that are carried out during the mining of the text and the extraction of insights from it. To carry out their respective tasks, these text mining strategies often make use of a variety of text mining tools and apps. Now, let us have a look at the various text mining approaches that are available: Let us now have a look at the text mining approaches that are considered to be the most well-known:

1. Information Extraction

This text mining method is by far the most well-known one. The process of gleaning useful information from massive swaths of textual material is referred to as information exchange. This method of text mining concentrates on figuring out how to extract entities, properties and the relationships between them from texts that are either semi-structured or unstructured. As information is extracted, it is then saved in a database so that it may be retrieved and accessed at a later time. The methods of accuracy and recall are utilised in order to assess and analyse the results for their usefulness as well as their relevance. The method known as "Information Extraction" is one that might be helpful when conducting analysis on textual material.

2. Information Retrieval

Information Retrieval, often known as IR, is the process of determining relevant and related patterns based on a certain word or phrase combination. Text mining is a technique that involves the application of various algorithms by IR systems in order to follow and analyse user actions and find relevant material in accordance with such behaviours. The search engines operated by Google and Yahoo are now the two most well-known IR systems. Information retrieval is the text mining technology that has gained the greatest notoriety throughout the years.

3. Categorization

This is one of the text mining strategies that is a form of "supervised" learning. In this type of learning, regular language texts are allocated to a specified set of themes

Notes

based upon the material they contain. Thus, categorization, or more accurately Natural Language Processing (NLP), is a procedure that involves collecting text documents, then processing and analysing those texts in order to discover the appropriate subjects or indexes for each document. In natural language processing (NLP), the co-referencing approach is a technique that is frequently used to derive meaningful synonyms and abbreviations from textual input. Currently, natural language processing has evolved into an automated process that can be utilised in a variety of scenarios, including the delivery of targeted ads, the filtering of spam, the classification of web pages according to hierarchical definitions and many other applications. Categorization is a handy method for doing analysis on the textual material that has been collected.

4. Clustering

The process of clustering is one of the most important strategies utilised in text mining. It aims to detect fundamental structures within textual material and arrange that information into relevant subgroups or “clusters” so that additional analysis may be performed on it. The development of meaningful clusters from the unlabeled textual data during the clustering process is a substantial problem because there is no previous information available on the textual data. Cluster analysis is a common text mining method that may either aid in the distribution of data or perform the function of a pre-processing step for other text mining algorithms that are executed on recognised clusters. Clustering is the most well-known approach that may be utilised in text mining.

5. Summarisation

The act of automatically creating a condensed version of a certain text that contains useful information for the end-user is referred to as “text summarisation.” Text mining is a process that involves reading through a number of different text sources in order to create summaries of longer texts that contain a considerable amount of information in a condensed format. The goal of this technique is to accomplish this while preserving the primary meaning and purpose of the original documents as much as possible. Text summarization is a strategy that integrates and combines the many different approaches that are used for text classification. Some examples of these approaches are decision trees, neural networks, regression models and swarm intelligence.

5.1.2 Definition and Language for Data Science

Data science is an interdisciplinary area that examines enormous volumes of data in order to discover hidden patterns, produce new insights and guide decision-making. This is accomplished via the use of algorithms, procedures and processes. Data scientists employ sophisticated machine learning algorithms to sift through, organise and learn from both organised and unstructured data in order to construct prediction models.

Data science is a rapidly developing topic that is applicable to a wide range of business sectors; hence, it offers a diverse range of employment prospects, ranging from computing to research. You will get an understanding of the applications of data science in the “real world,” as well as the career outlook for the subject, the requisite skills and the certifications that are necessary to secure employment in the area.

Data science definition

The study of data is referred to as data science. Data science is analogous to marine biology, which is the study of marine-dwelling organic life forms. Data scientists generate questions based on certain data sets and then utilise data analytics and advanced analytics to look for trends, develop prediction models and come up with insights that help organisations make decisions.

What is Data Science used for?

- **Descriptive Analysis**

It assists in showing data points properly for patterns that may arise that fulfil all of the restrictions that are imposed by the data. In other words, it entails organising, sorting and altering data in order to generate information that provides information that is insightful about the data that is presented. In addition to this, it requires transforming the raw data into a format that can be easily comprehended and interpreted by the user.

- **Predictive Analysis**

The practise of forecasting future results by using previous data in conjunction with a variety of methods like as data mining, statistical modelling and machine learning is known as predictive analytics. Businesses utilise predictive analytics to identify potential threats and opportunities by analysing patterns in the aforementioned data.

- **Diagnostic Analysis**

An in-depth investigation of the circumstances around an event is what this term refers to. Methods such as drill-down, data discovery, data mining and correlations are utilised in the process of describing it. For every given data collection, a variety of data operations and transformations may be carried out in order to search for and identify one-of-a-kind patterns using any of these methods.

- **Prescriptive Analysis**

The utilisation of predicted data is further developed through the use of prescriptive analysis. It makes a prediction about what is most likely to take place and recommends the most effective strategy for coping with that outcome. It is able to evaluate the potential outcomes of a number of options and provide recommendations for the best way to proceed. Machine learning recommendation engines, complex event processing, neural networks, simulation, graph analysis and simulation are all utilised in this process.

What is the Data Science process?

- **Obtaining the data**

The first thing that has to be done is to figure out what kind of data needs to be examined and then you need to export that data into either an Excel file or a CSV file.

- **Scrubbing the data**

It is necessary because before you can read the data, you need to make sure that it is in a state that is completely legible, without any errors and without any missing or incorrect information.

Notes

- **Exploratory Analysis**

Visualizing the data in a variety of different ways and recognising trends in order to search for anything that is not typical are both necessary steps in the analysis process. In order to assess the data, you need to have a very good attention to detail so that you can spot anything that is incorrect or missing.

- **Modeling or Machine Learning**

Based on the data that has to be processed, a data engineer or scientist will lay down instructions for the Machine Learning algorithm to follow in order to complete its task. The algorithm makes use of these instructions in an iterative manner in order to produce the desired output.

- **Interpreting the data**

At this point, you will reveal your results to the organisation and give a presentation on them. Your ability to communicate your results is going to be the single most important talent you possess in this situation.

What are different Data Science tools?

Here are a few examples of tools that will assist Data Scientists in making their job easier.

- Data Analysis – Informatica PowerCentre, Rapidminer, Excel, SAS
- Data Visualization – Tableau, Qlikview, RAW, Jupyter
- Data Warehousing – Apache Hadoop, Informatica/Talend, Microsoft HD insights
- Data Modelling – H2O.ai, Datarobot, Azure ML Studio, Mahout

Benefits of Data Science in Business

- Increases the ability to forecast business outcomes
- Analysis of a variety of complicated data
- Better decision making
- Development of new products
- Improves data security
- The creation of products with the end user in mind

Applications of Data Science

- **Product Recommendation**

Customers might be influenced to purchase comparable items through the use of the product suggestion strategy. For instance, a salesman at Big Bazaar is attempting to boost the store's revenue by offering discounts and combining things together in an effort to sell more of each item. As a result, he discounted the price of the shampoo and conditioner sets by bundling them together. In addition, clients will purchase both of them together at a price that is lowered.

- **Future Forecasting**

It is one of the methods that is utilised frequently in the field of Data Science. The

forecasting of both the weather and the future is carried out on the basis of a wide variety of data types that have been gathered from a wide variety of sources.

- **Fraud and Risk Detection**

This is one of the uses of data science that makes the most sense. Due to the proliferation of online transactions, it is possible for you to lose your data. For instance, the identification of fraudulent activity on credit cards is dependent on the amount, merchant, location and time of the transaction, among other factors. In the event that any of them seems odd, the transaction will be automatically invalidated and your card will be blocked for a period of at least 24 hours.

- **Self-Driving Car**

One of the most influential and influential innovations in the modern world is the self-driving automobile. We are teaching our automobile to reason and decide on its own based on the information it has accumulated. During this phase of the process, we have the ability to punish our model if it does not perform enough. When it begins gaining knowledge from all of its experiences in real time, the automobile gradually develops a higher level of intelligence over time.

- **Image Recognition**

Data science can detect the object in a picture and then classify it for you when you want to recognise some photographs. Face recognition is perhaps the most well-known use of image recognition technology. If you instruct your smartphone to unblock face recognition, it will scan your face. Hence, the system will initially identify the face, following which it will determine whether yours is a human face and last, it will determine whether or not the phone actually belongs to its rightful owner.

- **Speech to text Convert**

Voice recognition refers to the process of teaching a computer to comprehend human language naturally. Virtual assistants such as Alexa, Siri and Google Assistant are already extremely commonplace today.

- **Healthcare**

Data Science contributes to several subfields within the healthcare industry, including Medical Picture Analysis, the Research and Development of New Drugs, Genetics and Genomics and the Provision of Virtual Patient Assistance.

- **Search Engines**

Search engines such as Google, Yahoo, Bing and Ask, amongst others, offer us with many results in a very short amount of time. Many different data science methods are responsible for making this a reality.

Data Science Programming Languages

Data scientists may make powerful use of computers in their work. They make it possible for us to handle, analyse and display our data sets in ways that would be physically impossible to accomplish otherwise. Take a look through the data science online course to learn everything there is to know about the subject of data science. Programming is a vital skill in data science, but there are many different programming languages available. The question now is, which language is necessary for data

Notes

science? The following is a list of the nine most important programming languages that data scientists should be familiar with:

1. Python

Python is a programming language that may be utilised for the development of any type of software due to its general-purpose nature. It is widely considered to be one of the best programming languages for use in data science. Python is well-known for having an easy-to-read syntax, as well as portability and readability of its code. Also, it operates on all of the major platforms, which contributes to its popularity among software developers. Python is a programming language that can be learned quickly and has a huge community of developers supporting it. As a result, there are many resources available to assist you in getting started with Python. It is also strong enough to be utilised by data scientists working in professional capacities.

Python is an excellent programming language for novices since it uses a straightforward form of the English language and gives a wide range of options for data structures. In addition to this, it is well known for being a language that can be understood by machines. If a student is going to be starting out in a firm as a fresher, then this language is the greatest choice for them to learn.

2. SQL (Structured Query Language)

The structured query language (SQL) is one of the most used computer languages in the world. You are able to build queries to extract information from your data sets using this declarative language that is used for communicating with databases. In addition, the language enables you to communicate with databases. Learning SQL early on in your data science career is recommended because it is a language that is utilised in virtually all sectors. SQL instructions may be performed in two different ways: interactively via a terminal window or through scripts that are embedded in other software applications like web browsers or word editors.

The field of data science makes use of a programming language called Structured Query, which is domain-specific in nature. SQL is used in data science to assist users in collecting data from databases and afterwards editing that data if the circumstance calls for it. Because of this, a student who aspires to work in the field of data science has to have a solid foundation in Structured Query Language and databases. One could wish to think about taking online courses to become a professional data scientist in order to achieve success in data science through the use of SQL.

3. R

R is a computer language for statistics that is frequently utilised for statistical analysis, data visualisation and several other types of data manipulation. R's user-friendliness and versatility in performing complicated analyses on massive datasets have contributed to the field of data science seeing a surge in its popularity in recent years. In addition, the fact that R language data science provides many packages for machine learning algorithms like linear regression, the k-nearest neighbour algorithm, random forest and neural networks, amongst others, makes it a popular choice among many businesses that are interested in integrating predictive analytics solutions into their business procedures. For instance, as of today, hundreds of packages have been

made available for R, which makes it possible to do analyses of financial markets and readily predict weather patterns!

Notes

4. Julia

Julia is an essential language for data science that aspires to be straightforward while retaining a high level of capability and has a syntax that is comparable to that of MATLAB or R. Users are able to test their code in a hurry because to Julia's interactive shell, which frees them from the need to concurrently type down their whole projects. In addition, it is quick and frugal with memory, which makes it an excellent choice for working with massive datasets. Because of this, writing code is more quicker and easier to understand since it enables you to concentrate just on the issue at hand rather than on making type declarations.

5. JavaScript

The creation of online apps and websites often takes place in the computer language known as JavaScript. Since then, it has evolved into the most widely used language for developing client-side programmes for use online. In addition to this, JavaScript is well-known for its adaptability, since it can be utilised for everything from straightforward animations to in-depth applications requiring artificial intelligence. Continue reading if you want to learn more about the coding languages used in data science.

6. Scala

Scala has quickly risen to become one of the most popular languages for use cases involving artificial intelligence and data science. Scala is generally regarded a hybrid language that may be used for data science between object-oriented languages such as Java and functional languages such as Haskell or Lisp. This is due to the fact that Scala is statically typed and is an object-oriented programming language. In addition, Scala offers numerous advantages, such as functional programming, concurrency and high performance, that make it an appealing option for data scientists to use as their programming language of choice.

7. Java

Java is a concurrent computer programming language that is class-based, object-oriented and was developed primarily to have as few implementation dependencies as possible. Java is a general-purpose programming language for computers. As a direct consequence of this, Java is the most suitable programming language for data science. It is designed to allow application developers to "write once, run anywhere" (WORA), which means that generated Java code may run on any platforms that support the Java virtual machine (JVM) or JavaScript engines. This is one of the goals of this technology. Yet, code that depends on platform-dependent capabilities may not function on all JVMs since those features are optional for the JVMs and are not needed to be implemented. To become a data scientist, you will need to become proficient in all of these data science coding languages.

Notes

5.1.3 Collection of Data-Hunting, Logging, Scrapping

Hunting for potential threats includes conducting proactive investigations within a network to seek for irregularities that might point to a security breach. It is a laborious and time-consuming procedure since there is a massive quantity of data that needs to be collected and evaluated and the pace at which this process is carried out can have an impact on how effectively it is carried out. On the other hand, by utilising appropriate procedures for data gathering and analysis, that situation may be vastly improved. The data fertility of an environment is one of the foundations of a good threat hunting programme. To put it another way, first and foremost, a company has to have an enterprise security system that is capable of gathering data. The data that was gleaned from it is quite helpful to those who are searching for potential dangers.

Enterprise security can benefit from the addition of human intelligence through the use of cyber threat hunters as a supplement to automated solutions. They are trained specialists in the field of information technology security who hunt out, record, keep an eye on and eliminate any dangers before they may create significant issues. In an ideal situation, they are security analysts that work for a company's IT department and are well familiar with the company's activities. But, in certain cases, they are an outside analyst.

The practise of threat hunting involves investigating unknown environmental factors. It goes beyond the capabilities of conventional detection systems such as security information and event management (SIEM), endpoint detection and response (EDR) and others. The data pertaining to security is combed through by threat hunters. They hunt for hidden malware or attackers and check for patterns of suspicious behaviour that a computer could have missed or determined to be addressed but is not actually resolved. They also look for patterns of suspicious activity that a computer might have missed. They also assist in patching an organisation's security system to avoid the same kind of cyberattack from happening again in the future.

What Kind of Data Are We Collecting?

In order to effectively fulfil your role as a danger hunter, you need access to sufficient data. You will not be able to hunt if you do not have the correct info. Let us take a look at the criteria that determine what kind of information should be employed for hunting. It is essential to keep in mind that identifying the appropriate data is contingent upon the specific information that will be the focus of your search. In general, data may be divided into the following three categories:

1. Endpoint Data

The endpoint data is generated by the endpoint devices that are located within the network. End-user devices like mobile phones, laptops and desktop computers are examples of the types of devices that fall under this category. Nevertheless, the term "devices" can also refer to hardware like servers (like in a data centre). The real meaning of the term "endpoint" can be defined in a number of different ways, but in most cases, it refers to the things that we have outlined above.

You will find it useful to capture the following information from within endpoints:

- **Process execution metadata:** This data will include information on the many processes that are active on hosts (endpoints). The metadata that will be the

most sought after will comprise command-line commands and arguments, as well as the names and IDs of process files.

- **Registry access data:** This data will be connected to registry objects, including key and value information, on endpoints that are based on the Windows operating system.
- **File data:** This data will include, for instance, the dates on which files on the host were created or edited. It will also include the files' sizes, types and the locations on the disc where they are kept.
- **Network data:** This information will be used to determine which process is the parent for network connections.
- **File prevalence:** The information shown here will offer light on the extent to which a file is present in the environment (host).

2. Network Data

This data will originate from many network devices, including firewalls, switches, routers, proxy servers and DNS servers, among other things. Most of your focus should be placed on obtaining the following information from network devices:

- **Network session data:** The information on connections between hosts on the network will be of particular importance here. This information will, for example, contain the source IP address and the destination IP address, the length periods of the connection (including the start and finish timings), netflow, IPFIX and any other comparable data sources.
- **Monitoring tool logs:** Network monitoring tools will gather connection-based flow statistics as well as application information. The data that has been logged is what you should be gathering at this point. Moreover, application metadata concerning HTTP, DNS and SMTP will be of importance.
- **Proxy logs:** In this section, you will be collecting HTTP data that contains information on outbound Web requests, such as internet resources that are being accessed within the internal network.
- **Domain Name System (DNS) Logs:** The logs that you will acquire from this section will contain data relating to the resolution of domain names. This will comprise mappings from domain names to IP addresses, as well as the identity of internal customers who are submitting resolution requests.
- **Firewall logs:** Data from the firewall logs is one of the most significant types of information that you will be gathering. At the edge of the network, it will have information about the traffic going through the network.
- **Switch and router logs:** The information included in these will, for the most part, reveal what is occurring behind your network.

3. Security Data

This data will come from many security devices and solutions such as SIEM, IPS and IDS systems as its primary sources. You should be gathering the information listed below from the various security solutions:

- **Threat intelligence:** This type of data will contain the indications as well as the tactics, methods and procedures (TTPs) that hostile entities are carrying

Notes

out on the network. It will also include the activities that these entities are carrying out.

- **Alerts:** The data in this section will contain notifications from systems like IDS and SIEMs, which will indicate that a ruleset was broken or that any other incident took place.
- **Friendly intelligence:** This data will for example comprise information on key assets, accepted organisation assets, personnel information and business procedures. The significance of these data lies in the fact that they will assist the hunter and the analyst in better comprehending the environment in which they work.

Types of threat hunting

The first step in the hunting process is formulating a hypothesis based on security data or a trigger. Both the hypothesis and the trigger are used as jumping off points for in-depth research of the possible dangers. And these more in-depth inquiries can take a variety of forms, including systematic, unstructured and situational hunting.

• Structured hunting

An indication of attack (IoA) and the tactics, methods and procedures (TTPs) of an attacker form the basis of a structured hunt. Every hunt is coordinated with and based on the tactics, techniques and procedures (TTP) of the threat actors. As a result, the hunter is typically able to recognise a danger actor even before the attacker has had the chance to inflict damage to the environment. This hunting type makes use of the MITRE Adversary Tactics Techniques and Common Knowledge (ATT&CK) framework, making use of both PRE-ATT&CK and enterprise frameworks. The connection to this framework can be found outside of the IBM website.

• Unstructured hunting

A trigger, which is one of several signs of compromise, serves as the impetus for the start of an unstructured quest (IoC). This trigger often alerts a hunter to check for patterns both before and after the spotting of an animal. The hunter can explore as far into the past as the data retention and previously linked infractions will allow, using this information to guide their approach.

• Situational or entity driven

An enterprise's internal risk assessment or a trends and vulnerabilities study particular to that enterprise's IT environment are the sources from which a situational hypothesis is derived. When analysed, crowd-sourced attack data reveals the most recent tactics, techniques and procedures used by active cyberthreats. This data is the source of entity-oriented leads. The hunter of threats might then conduct a search inside the surroundings to look for these particular behaviours.

Hunting Models

a) Intel-based hunting

Intel-based hunting is a reactive hunting approach (link lives outside of ibm.com) that employs indicators of compromise derived from several sources of threat intelligence. The quest then proceeds in accordance with the predetermined rules that have been created by the SIEM and threat intelligence.

Intelligence-based hunts have the ability to make use of indicators of compromise, hash values, IP addresses, domain names, networks, or host artefacts that are given by intelligence sharing systems such as computer emergency response teams (CERT). It is possible to export an automated alert from these platforms and then feed it into the security information and event management system (SIEM) in the form of structured threat information expression (STIX) (link resides outside of ibm.com) and trusted automated exchange of intelligence information (TAXII) (link resides outside of ibm.com). After the SIEM has generated an alert based on an IoC, the threat hunter is able to analyse the malicious activities that occurred before and after the alert to determine whether or not the environment has been compromised.

b) Hypothesis hunting

The employment of a threat hunting library is part of the proactive hunting methodology known as “hypothesis hunting.” It identifies advanced persistent threat groups and malware assaults by using global detection playbooks and is connected with the MITRE ATT&CK paradigm.

The IoAs and TTPs of the adversary are utilised in hypothesis-based hunts. In order to provide a hypothesis that is in line with the MITRE framework, the hunter determines which threat actors are present based on the environment, domain and attack behaviours that are used. The threat hunter will monitor activity patterns once a behaviour has been detected in order to detect, identify and eventually isolate the danger. In this manner, the hunter is able to discover threat actors in a proactive manner before they are able to cause damage to an environment.

c) Custom hunting

The practise of custom hunting is predicated on situational awareness and hunting approaches that are industry-based. It detects anomalies in the SIEM and EDR tools and may be customised to meet the specific requirements of each individual customer.

Bespoke hunts, also known as situational hunts, are carried out in response to the specific requirements of individual clients or are carried out proactively in response to specific conditions, such as geopolitical difficulties or targeted assaults. These hunting endeavours are able to make use of both intelligence- and hypothesis-based hunting models, making use of information obtained via IoA and IoC.

• Data Logging

The practise of recording, storing and presenting one or more datasets for the purpose of doing activity analysis, locating trends and assisting in the prediction of future occurrences is known as data logging. Data logging can be done manually, but most operations are automated using intelligent programmes such as artificial intelligence (AI), machine learning (ML), or robotic process automation. Data recording can be done manually (RPA).

Data loggers have many applications across many different sectors, including the monitoring of supply chain activity and transportation activity, the measurement of temperature and humidity levels in a variety of locations, the monitoring of growing conditions and environmental conditions in greenhouses or farms and the review of network performance and CPU usage.

Notes

How does data logging work?

The process of data logging may be broken down into four primary steps:

1. A sensor is a device that collects and stores the data from one or more sources.
2. After this, a microprocessor will carry out fundamental measurements and logical operations, such as adding, subtracting, transferring and comparing numerical values.
3. The information that has been logged and saved in the memory unit of the data logger is then transmitted to a computer or another electronic device so that it may be analysed.
4. Upon the completion of the analysis, the data is then presented in the form of a knowledge graph or chart.

Four types of data loggers

Data loggers fall into four basic categories:

1. Standalone data logger
2. Wireless data logger
3. Computer-based data logger
4. Web-based data logger

Standalone data loggers

Standalone data loggers, also known as standalone sensors, are often very compact and portable electronic devices that have a USB interface. These devices each have the capability of having either an internal or an external sensor, which gives the device the ability to track data from either an on-site or a remote location, respectively.

Wireless data loggers

A standalone data logger called a wireless logger, also known as a wireless sensor, is a form of data logger that retrieves data through the use of wireless technology (such as a mobile app or Bluetooth) and then sends that data using cloud technology. Because of this, there is no longer a requirement to manually get and compile data from a variety of systems.

When compared to a standalone sensor, the speed of data collection is the primary advantage of utilising a wireless data logger. Cloud computing services have the potential to make it possible for a system to automate the transfer of data at regular or consistent periods. In practise, the approach is far more efficient than the traditional method of manually downloading data from a sensor.

Computer-based data loggers

Computer-based data loggers, also known as computer-based sensors, are data loggers that are physically connected to a computer. This is indicated by the name of the device. Real-time visibility into sensor data can be supported by a logger that is based on a computer and real-time analysis can be enabled by software programmes that run on

the computer. The fact that it is constrained to function only on certain operating systems is the most significant disadvantage of a logger that is based on a computer.

Notes

Web-based data loggers

The most cutting-edge variety of data loggers are known as web-based data loggers, sometimes known as web-based sensors. This computer is linked to the internet, which is accomplished most of the time by means of a wireless network; nevertheless, an ethernet connection may still be utilised in some circumstances. The collected data is then sent to a distant server, where it is both stored and made available on demand.

Web-based sensors, much like computer-based data loggers, have the potential to provide real-time monitoring and analysis. A computer-based sensor, on the other hand, has the ability to provide real-time warnings based on the recording levels that have been predetermined by the IT team. This feature can be beneficial to the company; nevertheless, it demands a large increase in the amount of energy that the logger produces. Because of this, the logger either needs its own power source or it may be prone to exhausting the battery of the endpoint with which it is attached. On the other hand, in contrast to computer-based loggers, web-based loggers are not constrained in terms of the operating system that the sensor may operate on.

• Data Scrapping

The act of importing data from websites into files or spreadsheets is referred to as “web scraping” and is also known as “data scraping.” It is utilised to take data from the web, either for the scraping operator’s own personal use or for the purpose of reusing the data on other websites. There is a wide variety of tools available that can automate the process of data scraping.

Scraping data is often used for a variety of purposes, including:

- Collecting business insight to inform site content.
- Calculate rates for sites that facilitate trip bookings or pricing comparisons.
- Use publicly available sources of data to either find potential customers or carry out market research.
- Provide information about products sold on eCommerce websites to third-party online marketplaces like Google Shopping.

The practise of “data scraping” has certain appropriate applications but is frequently misused by malicious individuals. One common application of data scraping is the collection of email addresses for the aim of sending spam or engaging in other fraudulent activities. The material of a website that is protected by copyright can also be scraped and automatically published on another website if scraping is utilised in this way. It is illegal in certain countries to utilise automated email harvesting tactics for the purpose of making a profit and the practise is widely seen as being immoral when it comes to marketing.

Data Scraping Techniques

The following are some of the most prevalent approaches that are used to scrape data from websites. In a nutshell, the process of web scraping involves retrieving

Notes

material from websites, processing that content with a scraping engine and creating one or more data files that include the content that was collected.

HTML Parsing

JavaScript is required for the process of parsing HTML, which can target either a linear or nested HTML page. It is a strong and quick way for scraping screens and retrieving resources, as well as collecting text and links (such as a nested link or email address, for example).

DOM Parsing

An XML file's structure, style and content are all defined by something called the Document Object Model (DOM). In order to gain a comprehensive understanding of the structure of online pages, scrapers often make use of DOM parsers. XPath and other tools may be used to scrape information from a web page using DOM parsers to get access to the nodes on the page that hold the information. Scrapers may extract full web pages by embedding web browsers such as Firefox and Internet Explorer, which allows them to process dynamically created material (or parts of them).

Vertical Aggregation

Platforms for vertical aggregation can be created to target certain verticals by businesses who have access to a significant amount of computer power. These are data harvesting platforms that are able to be deployed on the cloud. They are used to automatically build and monitor bots for certain verticals with minimum involvement from humans. Bots are developed in accordance with the information that is necessary for each vertical and the quality of the data that they extract is what determines how effective they are.

XPath

XPath is an abbreviation for XML Path Language, which is a query language for XML documents. XPath is an acronym for this language. As XML documents are organised in a tree-like form, scrapers may utilise XPath to browse through XML documents by picking nodes based on a variety of criteria. A scraper could use DOM parsing in conjunction with XPath in order to harvest whole web pages and then publish them on a target website.

Google Sheets

One of the most common tools for data scraping is Google Sheets, which includes a function called IMPORTXML that can be used to scrape data from a website. This is helpful for scrapers that wish to extract a certain pattern or data from the website. With this command, it is also possible to verify whether or not a website is secured and whether or not it may be scraped.

5.1.4 Cleaning Data-Artifacts, Data Compatibility

- **Data Cleaning**

Data cleaning is the process of fixing or removing wrong, corrupted, incorrectly formatted, duplicate, or incomplete data from a dataset. There are many ways for data

to be duplicated or mislabelled when you combine data from different sources. If the data is wrong, the results and algorithms are also wrong, even if they look right. There is no one way to say exactly what steps need to be taken to clean up data, because the steps change from dataset to dataset. But it's important to set up a template for your data cleaning process so you can be sure you're always doing it right.

How to clean data

Step 1: Remove duplicate or irrelevant observations

Take out any observations that do not belong, like duplicates or observations that don't matter. Most of the time, duplicate observations will happen when getting data. When you combine data from different places, "scrape" data, or get data from clients or different departments, you might end up with duplicate data. In this process, de-duplication is one of the most important things to think about. When you make observations that have nothing to do with the problem you are trying to solve, you have made irrelevant observations. For example, if you want to look at customer data about millennials but your dataset also has information about older generations, you might get rid of the older observations. This can make analysis faster and less likely to get in the way of your main goal. It can also make your dataset easier to work with and more effective.

Step 2: Fix structural errors

When you measure or move data and find strange naming conventions, typos, or wrong capitalization, you have made a structural error. These differences can lead to categories or classes that have the wrong names. For instance, you might see both "N/A" and "Not Applicable," but they should be looked at as the same category.

Step 3: Filter unwanted outliers

There will often be one-off observations that do not seem to fit with the rest of the data you are looking at. If you have a good reason to get rid of an outlier, like bad data entry, that will help the data you are working with work better. But the appearance of an outlier can sometimes prove a theory you are working on. Remember that just because there is an outlier doesn't mean it is wrong. This step is needed to figure out if that number is correct. If an outlier turns out to be useless for analysis or a mistake, you might want to get rid of it.

Step 4: Handle missing data

You cannot just ignore missing data because many algorithms can't handle them. There are a few ways to deal with data that is not there. Neither is best, but both can be thought about.

1. You can drop observations that have missing values as a first option, but this will cause information to be lost, so be aware of this before you do it.
2. The second option is to fill in missing values based on other observations. Again, there is a chance that the data integrity will be lost because you may be using assumptions instead of real observations.

Notes

3. As a third option, you could change how the data is used to navigate null values more effectively.

Step 5: Validate and QA

As part of basic validation, you should be able to answer these questions at the end of the data cleaning process:

- Does the data make sense?
- Does the information follow the rules for the field?
- Does it show that your working theory is right or wrong, or does it give you any new information?
- Can you find patterns in the data that will help you come up with your next idea?
- If not, is it because of a problem with the data?

Bad business decisions and strategies can be based on wrong or “dirty” data that leads to false conclusions. False conclusions can make you look bad in a reporting meeting when you find out that your data does not hold up. Before you get there, you need to make sure that your organisation has a culture of good data. To do this, you should write down what data quality means to you and what tools you might use to create this culture.

• Data Artifact

A fault in the data that is created by the apparatus, the procedures, or the environment is called an artefact. Errors in hardware or software, situations such as electromagnetic interference and poor designs such as an algorithm prone to miscalculations are common causes of data defects. Other sources of data faults include factors such as electromagnetic interference.

An artefact is a result of software development that helps characterise the architecture, design and operation of software. Artifacts are often referred to as software artefacts. Artifacts are similar to road maps in that developers of software may use them to track every step of the software creation process.

Databases, data models, written texts and scripts are all examples of possible artefacts. Since developers may utilise artefacts as reference material to assist in problem resolution, they are beneficial to the process of maintaining and upgrading software. Artifacts are given documentation and placed in a repository in order for software developers to be able to access them whenever they are needed.

Throughout the course of the software development process, artefacts of the programme are often formed. These artefacts may relate to certain procedures or procedures involved in the creation of the software. A software build, for instance, includes the programmer's code in addition to a variety of artefacts from the software's development. The operation of the programme is made possible by some of these artefacts, while others of them provide an explanation of how the software operates. For instance, the code's artefacts may consist of a list of dependencies, the source code for the project, or a collection of resources. These artefacts are retained in a repository so that they may continue to be organised and can be accessed whenever there is a need to do so.

When they have been produced, artefacts are essential in all stages of the software development process. The process of producing software is made easier with the aid of artefacts created specifically for that purpose. In the event that an artefact that defines the architecture, design and function of a piece of software is absent, it is possible that this will leave developers in the dark in the event that something goes wrong. The ability for developers to access artefacts at any time and from a single location is made possible by storing relevant artefacts in a repository.

The operation and functionality of a piece of software is characterised by its artefacts, which may include control sequences or database queries. Developers are able to comprehend how software operates with the assistance of artefacts, as opposed to having to examine the intricate coding that lies behind it. This is particularly helpful for developers who have just been brought on board, since the artefacts enable them to comprehend the thinking process of developers who came before them. While running the programme, performing maintenance on it, or updating it, it is helpful to be able to look at artefacts that provide a concise explanation of how the product works.

Types of software artifacts

The following is a list of the three primary classifications that artefacts can be placed into:

- **Code-related artifacts.** This code serves as the basis for the software and provides the programmer with the ability to test the product prior to releasing it to the public. The compiled code, the setup scripts, the test suites, the created objects and the logs generated during testing and quality assurance can all be considered code artefacts.
- **Project management artifacts.** When the code has been constructed, these artefacts are produced so that its functioning can be evaluated. The minimum necessary standards, benchmarks, project vision statements, roadmaps, change logs, scope management plans and quality plans are all examples of artefacts that fall under this category.
- **Documentation artifacts.** These artefacts are responsible for maintaining a record of pertinent papers such as schematics, end-user agreements, internal documentation and written manuals.

Data Manipulation

The process of arranging data in such a way that it is simpler to understand, more designed, or more organised is referred to as “data manipulation.” For the sake of clarity, a compilation of any form of material may be arranged in alphabetical order, for instance. This would make it much simpler to comprehend. On the other hand, if the information pertaining to all of the employees in a company is not structured, it may be difficult to discover information on a specific individual working for the firm. Consequently, it is possible that all of the employees’ information may be grouped in alphabetical order, which would make it much simpler to obtain information about any specific employee. The proprietors of websites can analyse their traffic sources and identify their most popular pages with the use of data manipulation. As a result, it is utilised rather commonly in web server logs.

Notes

Users in accounting and other disciplines related to it also make use of data manipulation in order to arrange data in order to calculate product costs, future tax responsibilities, pricing trends and other similar things. In addition to this, it assists those who make predictions about the stock market to foresee trends and determine how stocks may perform in the near future. In addition, computers may utilise data manipulation to show information to consumers in a manner that is more realistic by basing it on web pages, the code in software programmes, or data formatting. This can be accomplished by manipulating the data.

The Data Manipulation Language, or DML, is a computer language that is used to modify data. The abbreviation stands for “Data Manipulation Language,” which is a programming language that facilitates the addition, deletion and modification of data as well as databases. It involves altering the material so that it may be read in a manner that is not difficult.

Objective of Data Manipulation

The manipulation of data is an essential component for the successful operation and optimisation of a business. You need to handle data in the appropriate manner and change it in order to transform it into information that has significance. For example, you may analyse trends, financial data, or customer behaviour. The manipulation of data provides an organisation with a number of benefits, some of which are outlined below:

- **Consistent data:** The process of data manipulation offers a technique to arrange your data in an inconsistent manner and transform it into a structured one that is simpler to read and more readily comprehended. When you collect data from a variety of sources, you might not have a unified perspective of the data; nevertheless, data manipulation ensures that the data is well-organised, formatted and stored in a consistent manner.
- **Project data:** Data manipulation is more valuable since it helps to give more in-depth research by using previous data to predict the future. This is especially true when it comes to money, where data manipulation is very useful.
- **Erase or ignore redundant data:** Data manipulation can help you manage your data and erase useless data that is constantly there. ○ Delete or ignore redundant data.

In general, you are able to do a wide variety of actions on the data, including edit, remove, update, convert and incorporate the data into a database. It contributes to the creation of additional value from the data. It is useless information if you do not have the skills necessary to utilise it in an efficient manner. Hence, being able to arrange your data in the appropriate manner will allow you to make better business decisions, which will be to your benefit.

Steps involved in Data Manipulation

You will find a list of critical actions that you should follow below, which should assist you in getting started with data manipulation.

1. To begin, in order to manipulate data, you must possess the data in the first place. As a result, you are obligated to develop a database that is constructed from several sources of data.

2. This knowledge has to be restructured and reorganised, which might be accomplished through the use of data manipulation, which assists you in purifying the information you have.
3. After that, in order to begin working with data, you will need to import a database and then build it.
4. You are able to alter, delete, merge, or combine the information you have at your disposal with the assistance of data manipulation.
5. Finally, the process of data manipulation makes it much simpler to analyse the data.

5.1.5 Dealing with Missing Values, Outliers

- **Missing Value**

The values or data that are not saved (or are not available) for some variable/s in the provided dataset are referred to as “missing data.” This is the definition of missing data. The following is an example of some of the data that is missing from the Titanic dataset. You can see that some of the numbers in the columns labelled “Age” and “Cabin” are missing.

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	71.5		S
2	1	1	female	38	1	0	PC 17599	71.233	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Why Is Data Missing from the Dataset?

There is a wide variety of potential explanations for why some numbers are absent from the data. The method used to handle missing data is impacted by the circumstances surrounding the absence of data in the dataset. So, it is essential to comprehend the reasons why the data can be absent.

The following is a list of some of the reasons:

- Past data could get damaged owing to faulty upkeep.
- Some fields do not have observations recorded for them for a variety of different reasons. There is a possibility that the values were not recorded correctly owing to human error.
- The user did not intend to supply the values in any of their submissions.
- If a participant does not react to a certain item, this indicates that they choose not to.

Types of Missing Values

Technically, the missing values may be broken down into the following categories:

Notes

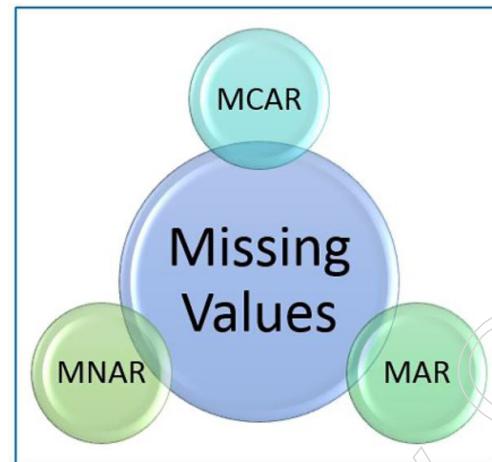


Figure 1 - Different Types of Missing Values in Datasets

Missing Completely at Random (MCAR)

In MCAR, the likelihood that an observation will be missing data is proportional to the total number of observations. In this instance, there is no connection between the missing data and any other values seen or unobserved (the data that is not recorded) within the supplied dataset. This is because the missing data do not correspond to any of the other values. That is to say, missing numbers are totally unrelated to the other data in any way. There is not a consistent trend.

In the case of the MCAR data, the value may be absent as a result of human mistake, the failure of some system or piece of equipment, the loss of a sample, or some undesirable technicalities that occurred when recording the results. Consider the situation where certain books at a library have been overdue for a certain amount of time. Inside the computer system, there are certain values of overdue books that are missing. It's possible that the cause was a mistake made by a person, like the librarian forgetting to key in the values. So, the values of overdue books that are missing have no connection to any of the other variables or data in the system. Due to the unusual nature of the situation, no assumptions should be made. The benefit of utilising such data is that the statistical analysis may continue to be objective.

Missing At Random (MAR)

MAR data indicates that the reason for missing values can be explained by variables on which you have complete information, as there is some relationship between the missing data and other values or data. This is because there is some kind of connection between the missing data and other values or data. In this instance, there are not any gaps in the data for any of the observations. It is only absent within specific subsamples of the data and the numbers that are missing follow a certain pattern overall.

For instance, if you review the data from the survey, you can discover that everyone has provided an answer to the question "Gender," but the majority of the responses to the "Age" question from persons who have identified themselves as "female" are missing. (The primary reason for this is that the majority of the women do not wish to disclose their ages.)

Hence, the only thing that affects the chance of data being absent is the value of data that has already been detected. In this particular instance, the variables "Gender" and "Age" are connected to one another. The 'Gender' variable can provide an explanation as to why there are missing values of the 'Age' variable, but you are unable to make a prediction regarding the missing value itself.

Let us say a survey is conducted on overdue books in a library. In the survey, questions on gender and the quantity of overdue books are posed. Suppose that the majority of the poll's respondents are female and that fewer men than women answered it. Thus, there must be another element at play and that component is gender. This explains why there is a gap in the data. In this particular scenario, the statistical analysis may produce biased results. Only by modelling the data that is lacking will one be able to obtain an estimate of the parameters that is objective.

Missing Not At Random (MNAR)

The missing numbers are dependent on the data that was not seen. If there is some structure or pattern in the missing data and other observable data cannot explain it, then the missing data is regarded to be missing not at random. [Case in point:] (MNAR).

MNAR is a possible classification for the missing data if neither MCAR nor MAR are applicable to the situation in question. It is possible for this to occur as a result of people's unwillingness to offer the necessary information. It is possible that a certain subset of responders to a survey won't answer some of the questions.

For illustration's sake, let us say the poll for a library asks for the name of the library as well as the quantity of books that are overdue. So, the vast majority of people who do not have any overdue books are likely to respond to the survey. Individuals who have more books that are overdue have a lower likelihood of responding to the survey. So, in this particular scenario, the missing value of the number of books that are overdue is dependent on the individuals who have a greater number of books that are overdue. One other illustration of this would be the fact that those with lower incomes are more likely to withhold certain information from a survey or questionnaire. Even in the case of MNAR, there is a possibility that the statistical analysis will be biased.

- **Outliners**

In the field of data analytics, outliers are values contained inside a dataset that differ dramatically from the rest in the set; typically, these values are either much higher or much lower than the others. Variabilities in a measurement, faults in an experiment, or even a new phenomenon might be indicated by outliers. In the actual world, the typical height of a giraffe is around five metres (or sixteen feet) tall. On the other hand, in recent times, researchers have found evidence of two giraffes that measure 9 and 8.5 feet in height, respectively. In comparison to the rest of the giraffe population, these two giraffes stand out as unusual cases.

Outliers are a potential source of abnormalities in the findings that are acquired after going through the process of data analysis. This indicates that they call for some more consideration and, in some instances, will need to be eliminated in order to carry out an accurate data analysis.

Notes

It is essential to the process of data analytics to focus extra attention on data points that are significantly different from the norm for two primary reasons:

1. The presence of outliers has the potential to provide an unfavourable outcome for an analysis.
2. An information that a data analyst needs from the study may be the behaviour of outliers, or the behaviour of outliers themselves.

Types of outliers

There are two kinds of outliers:

- An excessive result that is only associated with one variable is referred to as a univariate outlier. For instance, Sultan Kosen, who stands at a height of 8 feet, 2.8 inches, is the tallest guy who is still living today (251cm). Because height is the only variable at play here, this instance is what's known as a "univariate outlier" because it's such an extreme example of only one element.
- The combination of uncommon or extreme values for at least two different variables is what constitutes a multivariate outlier. If you were to look at the height and weight of a group of adults, for instance, you might notice that one of the people in your dataset has a height of 5 feet and 9 inches, which is a measurement that is considered to be within the normal range for the variable in question. Another example would be if you were to look at the ages of the people in your group. You could also notice that this guy has a weight of 230 kilogrammes (110 lb). Again, this observation by itself is consistent with values that are considered to be within the usual range for the variable of interest, which is weight. A person of this age who is 5 feet and 9 inches tall and weighs 110 pounds is an unexpected combination; but, when you examine these two facts together, you arrive at this conclusion. That's an extreme value in more than one variable.

In addition to the differentiation between univariate and multivariate outliers, you may also come across outliers classified as any of the following:

- **Global outliers**, often referred to as point outliers, are single data points that are located at a significant distance from the remainder of the data distribution.
- **Contextual outliers**, also known as conditional outliers, are values that considerably depart from the rest of the data points that are found in the same context. This indicates that the same value might not be deemed an outlier if it happened in a different context. With time series data, it is not uncommon to come across outliers that fall into this category.
- The term "**collective outliers**" refers to a collection of data points that are viewed as being fully unique in comparison to the overall dataset.

5.2 Big Data Fundamentals and Hadoop Integration with R

Introduction

Big Data refers to a collection of data that is not only enormous in amount but also expanding at an exponential rate over time. Because of its enormous quantity and high degree of complexity, none of the conventional methods for managing data are able

to store it or handle it in an effective manner. Big data is simply data that is stored in extremely large quantities.

5.2.1 Definition, Evolution of Big Data and its Importance

Big Data, a lately popular term, is described as a significant volume of data that cannot be stored or processed using standard data storage or processing technology. Because of the large volumes of data generated by human and machine operations, the data is so complicated and vast that it cannot be comprehended by humans or fit into a relational database for analysis. When properly analysed using current tools, however, these huge amounts of data present enterprises with important insights that help them enhance their business by making educated decisions.

Big Data is a term used to describe a collection of massive and complicated datasets that are challenging to analyse using older data processing systems.

Types of Big Data

Big Data is essentially classified into three types:

- Structured Data
- Unstructured Data
- Semi-structured Data

The three categories of Big Data mentioned above are technically relevant at all levels of analytics. When working with enormous amounts of big data, it is vital to understand the source of raw data and how it is treated before analysis. Because there is so much data, information extraction must be done effectively in order to get the most out of the data.

Structured Data

Structured data is the most orderly and consequently the most convenient to deal with. Its dimensions are determined by predefined specifications. Like spreadsheets, each item of information is organised into rows and columns. Structured data contains quantitative information such as age, contact information, address, billing information, costs, debit or credit card numbers and so on.

Since structured data is quantifiable, it is simple for programs to filter through and gather data. Processing structured data takes minimal to no preparation. The data merely has to be cleaned and reduced to the most important aspects. To conduct a proper investigation, the data does not need to be transformed or examined in great detail.

Structured data follows road maps to specified data points or schemas to outline the position and significance of each datum.

One of the benefits of structured data is the simplified process of combining corporate data with relational data. Because the relevant data dimensions have been established and are in a consistent format, relatively little preparation is necessary to ensure that all sources are compatible.

For structured data, the ETL (Extract, Transform, Load) process puts the end result in a data warehouse. The original data is acquired for a specific analytics objective and

Notes

the databases are heavily organised and filtered for this purpose. However, there is a limited quantity of structured data available and it constitutes a small percentage of all extant data. According to consensus, structured data accounts for about 20% or less of total data.

Unstructured Data

Not all data is well-structured and organised, with clear instructions on how to use it. Unstructured data refers to all disorganised data. Almost all data produced by a computer is unstructured data. It might take a long time and a lot of work to make unstructured data understandable. Datasets must be interpretable in order to provide meaningful value. However, the act of making it happen may be far more satisfying.

The most difficult aspect of unstructured data analysis is educating an application to comprehend the information it is retrieving. Often, translation into organised form is necessary, which is difficult and varies depending on the format and final purpose. Text parsing, NLP and building content hierarchies using taxonomy are some approaches for achieving translation. Complex algorithms are used to integrate the scanning, interpreting and contextualizing operations.

Unstructured data is stored in data lakes, as opposed to structured data, which is saved in data warehouses. Data lakes save both the raw format of data and all of its metadata. In contrast to data warehouses, where data is constrained to its predetermined format, data lakes make data more pliable.

Semi-structured Data

Semi-structured data is intermediate between structured and unstructured data. It typically refers to unstructured data with information attached. Semi-structured data, such as location, time, email address, or device ID stamp, can be inherited. It might even be a semantic tag that is later added to the data.

Consider the following example: an email. The time an email was sent, the sender's and recipient's email addresses, the IP address of the device from which the email was sent and other pertinent information are connected to the email's content. While the actual content is not organised, these components allow the data to be structurally categorised.

Semi-structured data may be transformed into a valuable asset by using the correct datasets. By linking patterns with metadata, it can help with machine learning and AI training. The lack of a fixed schema in semi-structured data can be both an advantage and a disadvantage. Putting forth all that effort to inform an application what each data item means might be difficult. At the same time, there are no definitional constraints in structured data ETL.

Evolution of Big Data

Since the early 1990s, the phrase "Big Data" has been in usage. John R. Mashey is credited with popularizing the phrase "Big Data." Big Data is not something that has just been used in the previous two decades. People have been attempting to employ data analysis and analytics approaches to aid in decision-making for many years. The massive rise in both organised and unstructured data sets made traditional data

analysis extremely challenging and this evolved into 'Big Data' in the previous decade. Big Data's evolution may be divided into three periods, each with its own set of features and capabilities that have led to the modern definition of Big Data.

Phase I: Big Data originated in the realm of database administration in Phase I. It is primarily determined by the storage, extraction and optimization of data contained in Relational Database Management Systems (RDBMS). In the initial phase of Big Data, the two key components are database administration and data warehousing. It lays the groundwork for current data analysis and approaches including database queries, online analytical processing and standard reporting tools.

Phase II: Beginning in the early 2000s, the use of the Internet and the World Wide Web began to provide novel data collecting and data analysis capabilities. Yahoo, Amazon and eBay expanded their online storefronts and began studying customer behaviour for customisation. HTTP-based online content significantly boosted semi-structured and unstructured data. Organisations now needed to develop new methodologies and storage solutions to cope with these new data kinds and successfully analyse them. The proliferation of social media data in following years exacerbated the need for tools, platforms and analytics methodologies capable of extracting valuable information from this unstructured data.

Phase III: Over the last decade, the widespread use of smart phones with various internet-based apps has enabled the analysis of behavioural data (such as clicks and search queries) as well as location-based data (GPS-data). Simultaneously, the proliferation of sensor-based internet-enabled gadgets known as the "Internet of Things" (IoT) is causing millions of TVs, thermostats, wearables and even refrigerators to create zettabytes of data every day. With the phenomenal expansion of 'Big Data,' a race to extract relevant and useful information from these new data sources has begun. This gives rise to other new phrases such as 'Big Data Analytics.'

Table 1 gives the summary of the three phases in Big Data.

Phase-I	Phase-II	Phase-III
DBMS-based, structured content: 1. RDBMS and data warehousing 2. Extract Transfer Load 3. Online Analytical Processing 4. Dashboards and scorecards 5. Data mining and statistical analysis	Web based, unstructured content 1. Infomiation retrieval and extraction 2. Opinion mining 3. Question answering 4. Web analytics and web intelligence 5. Social media analytics 6. Social network analysis 7. Spatial-temporal analysis	Mobile and senor-based content 1. Location-aware analysis 2. Person-centred analysis 3. Context-relevant analysis 4. Mobile visualization 5. Human-Computer interaction

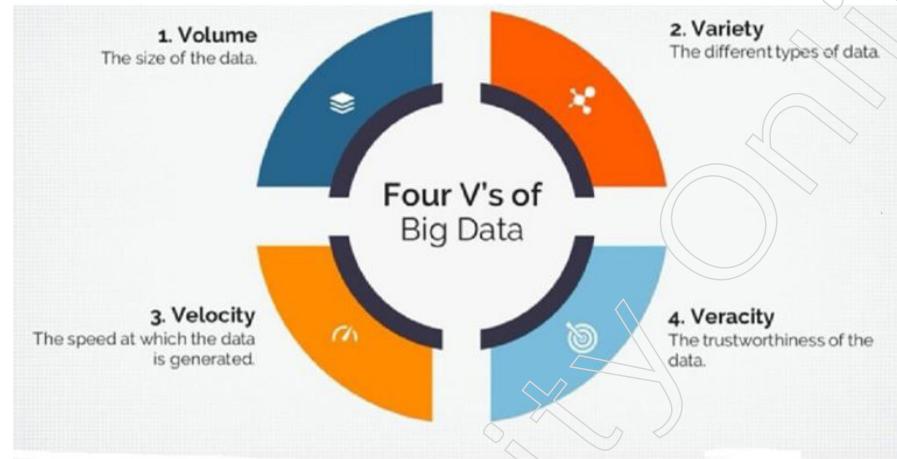
5.2.2 Four Vs in Big Data, Drivers for Big data

Big Data is made up of a significant volume of data that is not handled by typical data storage or processing units. Many global corporations utilize it to process the data

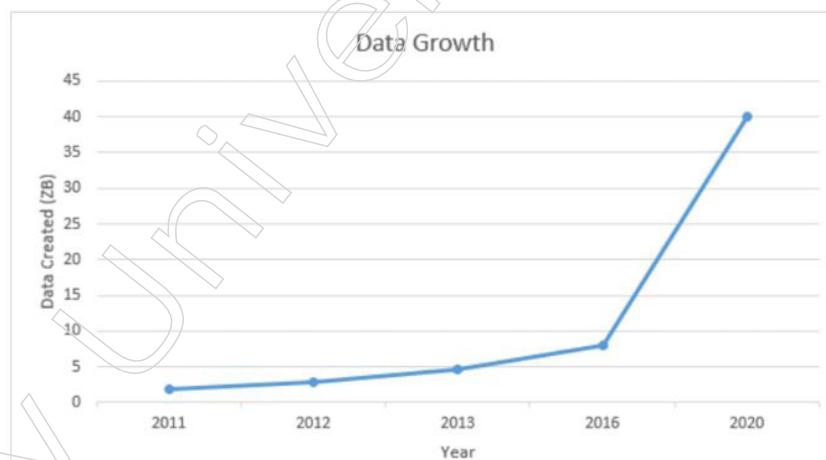
Notes

and business of numerous firms. Before replication, the data flow would reach 150 exabytes per day.

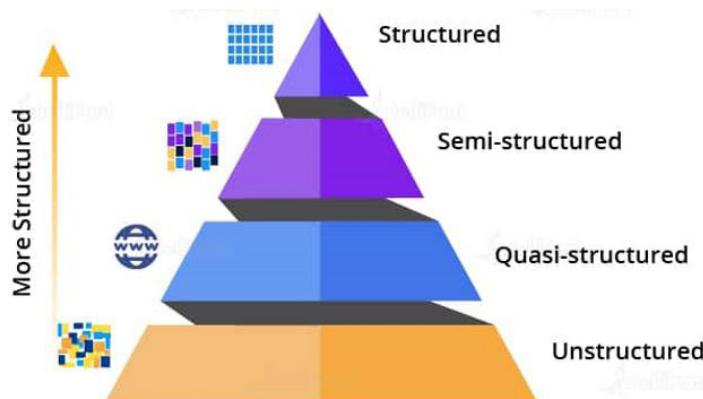
Big Data has four v's that explain its properties.



- Volume:** This refers to extremely enormous amounts of data. As the graphic shows, the volume of data is increasing at an exponential rate. The data generated in 2016 was just 8 ZB; by 2020, the data is predicted to be 40 ZB, which is extraordinarily huge.



- Variety:** One explanation for the rapid increase in data volume is that data is arriving from diverse sources in varied forms. We've already spoken about how data is classified into distinct categories. Let's take another look at it with some additional instances.

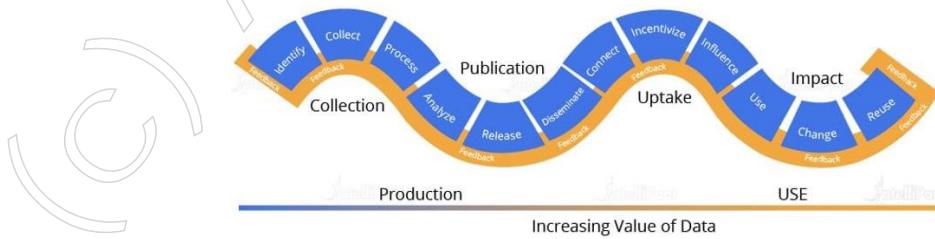


- a) **Structured Data:** Data is presented in an organised format with all needed columns. It is organised or tabular in nature. Structured data is data that is stored in a relational database management system. For example, the data in the employee table below, which is stored in a database, is structured.

Emp. ID	Emp. Name	Gender	Department	Salary (INR)
2383	ABC	Male	Finance	650,000
4623	XYZ	Male	Admin	5,000,000

- b) **Semi-structured Data:** The schema is not fully specified in this kind of data, hence both forms of data are present. As a result, semi-structured data has a structured form but is not specified, such as JSON, XML, CSV, TSV and email. Unstructured web application data includes transaction history files, log files and so on. Online Transaction Processing (OLTP) systems are designed to operate with structured data, which is kept in relations, often known as tables.
- c) **Unstructured Data:** All unstructured files, such as video, log, audio and picture files, are included in this data format. Unstructured data is any data that has an unknown model or organisation. Because of its huge quantity, unstructured data presents a number of processing issues when it comes to extracting value from it. A complicated data source including a mix of text files, videos and photos is one example of this. Several businesses have a lot of data, but they don't know how to extract value from it since the data is in its raw form.
- d) **Quasi-structured Data:** This data format consists of textual data with inconsistencies in data formats that may be formatted with work, time and the assistance of various technologies. Web server logs, for example, are a log file that is automatically produced and maintained by a server and provides a record of actions.

3. **Velocity:** The rate at which data accumulates determines whether the data is large data or regular data. The term velocity refers to the rapid collection of data. In Big Data velocity, data pours in from many sources such as machines, networks, social media, mobile phones and so on. A large and constant influx of data exists. This influences the data's potential, or how quickly it is created and processed to fulfill needs. Data sampling can aid in dealing with issues such as 'velocity.' Example: Google receives more than 3.5 billion searches every day. Furthermore, Facebook users are expanding by around 22% year on year.
4. **Value:** How will data extraction work? Our fourth V enters the picture here; it is concerned with a system for determining the right interpretation of data. First and first, you must mine data, which is the act of converting raw data into meaningful data. The data that you have cleaned or obtained from the raw data is next analysed. Then, you must ensure that whatever analysis you have performed benefits your business, such as discovering insights, outcomes and so on in ways that were not previously possible.



Notes

You must ensure that any raw data provided to you for the purpose of gaining business insights is cleaned up. After you've cleansed the data, a problem arises: certain packages may be lost during the process of dumping a big amount of data. So, in order to remedy this problem, our next V enters the scene.

Importance of Big Data

The significance of big data is not contingent on the quantity of data possessed by an organisation. The manner in which the organisation makes use of the information that it has obtained is directly related to its significance. Every organisation does things differently with the data that it has acquired. The more efficiently the firm makes use of its data, the quicker it will expand. The businesses competing in the modern market are required to amass and examine this information because:

1. Cost Savings

When it comes to storing big volumes of data, organisations may reap the benefits of Big Data technologies like Apache Hadoop and Spark, which help them save money in the process. These tools assist firms in determining business practises that are more productive and efficient.

2. Time-Saving

Companies are able to acquire data from a wide variety of sources with the assistance of real-time analytics that run in memory. They are able to swiftly examine data with the assistance of tools such as Hadoop, which enables them to make decisions more quickly that are founded on their discoveries.

3. Understand the market circumstances

The study of Big Data enables firms to have a better grasp of the current state of the market. For instance, doing a study of the purchase patterns of customers enables businesses to determine which items are the most popular and, as a result, to make more of those things. This allows businesses to get a competitive advantage over their other rivals.

4. Social Media Listening

With the technologies for big data, businesses are able to do sentiment analysis. These things make it possible for them to obtain comments about their firm, or information regarding who is saying what regarding the organisation. Tools for analysing large amounts of data can help businesses enhance their internet presence.

5. Improve Your Capacity to Acquire and Customer Retention

Clients are an essential resource that are necessary for the success of every organisation. Without establishing a solid foundation of loyal customers, no company can hope to attain lasting success. Nonetheless, despite having a stable consumer base, the businesses are unable to ignore the rivalry that exists in the market.

If we are unable to understand what it is that our clients desire, then the success of our businesses will suffer. That will lead to a decrease in the number of customers, which will have a negative impact on the expansion of the firm.

Analytics of large amounts of data help organisations recognise patterns and trends connected to their customers. Analysis of the behaviour of customers is the path to a successful business.

6. Solve Advertisers Problem and Offer Marketing Insights

The entirety of corporate operations is shaped by big data analytics. It gives businesses the ability to meet the requirements of their customers. The company's product range may be modified with the assistance of big data analytics. It makes certain that marketing initiatives are successful.

7. The driver of Innovations and Product Development

The availability of large amounts of data gives businesses the ability to develop new goods and improve existing ones.

5.2.3 Big Data Analytics, Big Data Applications

- **Big Data Analytics**

Big data analytics is a term that refers to the methodology, tools and applications that are used to collect, analyse and derive insights from a wide variety of data sets that move at a high volume and velocity. These data sets may have originated from a wide range of sources, such as the internet, mobile devices, email, social media platforms, or networked smart devices. They typically contain data that is generated at a quick speed and in a range of formats, ranging from organised (database tables, Excel sheets) to semi-structured (XML files, websites) to unstructured (text files). They can also contain data that is unorganised (images, audio files).

Because traditional data analysis software is unable to manage this level of complexity and bulk, the advent of systems, tools and applications designed specifically for the study of big data has given rise to the need for a solution.

Advantages of Big Data Analytics

1. Risk Management

Big Data Analytics gives vital insights into customer behaviour and market trends, enabling businesses to assess both their current standing and their potential for future development. They may also use predictive analytics to foresee the possibility of prospective threats and prescriptive analytics and other types of statistical analysis methodologies to reduce the likelihood of such dangers.

2. Product Development and Innovations

Big Data Analytics may also help businesses decide whether or not to produce a product based on how well it will sell in the market. The feedback of customers on a product is included in big data. The performance of a company's product is then evaluated using these statistics and the company decides whether or not production of the product should be continued or halted.

When it comes to creativity, the information that was gleaned from the survey is essential. They may be utilised to enhance a range of things, including business

Notes

strategies, marketing methods and a variety of other things. Because modern businesses increasingly rely on market insights to construct any kind of corporate plan, the number of occasions in which big data might be advantageous is practically limitless.

3. Quicker and Better Decision making

The pace at which decisions are made has quickened in tandem with the acceleration of globalisation. The decision-making process has been sped up because of big data analytics. Businesses are no longer required to wait for responses over the course of days or months. Efficiency has also increased as a direct result of the decreased reaction time. Because employing this technique allows businesses to restructure their business models, companies no longer have to take significant financial hits in the event that their product or service is not favourably received by consumers.

4. Enhance the Customer Experience

When companies are able to regularly monitor the behaviour of their customers, they may be able to provide a more personalised level of service to those customers. The use of diagnostic analytics may be helpful in locating solutions to the problems that the consumer is experiencing. Because of this, the customer will have a more personalised experience, which will ultimately lead to improved customer satisfaction.

5. Complex Supplier Networks

Big data allows businesses to improve the level of accuracy with which they serve B2B communities, often known as supplier networks. The use of Big Data analytics gives suppliers the ability to triumph over the limitations they experience. It opens the door for providers to make use of larger degrees of contextual intelligence, which ultimately contributes to an increase in their level of success.

6. Focused and Targeted Campaigns

Platforms may make advantage of big data in order to provide their target market with individualised products. instead of wasting money on advertising methods that are not working. The use of big data enables businesses to do more sophisticated analysis of customer trends. This entails analysing transactions made at physical stores as well as those made over the internet. These insights, in turn, allow businesses to build campaigns that are lucrative, specific and targeted, so helping them to fulfil the expectations of their customers and improve their loyalty to the brand.

• Big Data Applications

Large quantities of complicated and unprocessed data are what are meant when people talk about "Big Data." Data scientists, analytical modellers and other experts are able to analyse a vast number of transactional data thanks to the usage of Big Data in today's businesses. This makes doing business significantly more informative and makes it possible to make business choices. The major information technology industries of the 21st century are propelled by the lucrative and potent fuel that is big data. The usage of big data is becoming widespread across all facets of the commercial world. The application of Big Data will be the topic of discussion in this section.

Travel and Tourism**Notes**

Big Data is being utilised in the travel and tourist industry. We are able to estimate travel facility requirements at various sites, increase business through dynamic pricing and a great deal more as a result of this capability.

Financial and banking sector

The usage of big data technologies is widespread throughout the financial and banking industries. Big data analytics may assist banks and customers better understand consumer behaviour on the basis of investment patterns, shopping trends, incentive to invest and inputs that are derived from either personal or financial backgrounds.

Healthcare

With the assistance of predictive analytics, licenced medical experts and other health care workers, the use of big data has begun to make a significant contribution to the field of medicine. It is also capable of producing individualised medical treatment for single patients.

Notes**Telecommunication and media**

The most prominent industries making use of big data are those in the telecommunications and multimedia fields. Every day, zettabytes of new data are created and in order to manage such massive amounts of information, big data solutions are required.

Government and Military

High rates of technology usage were also observed in government and military institutions. On the record, we are privy to the statistics that the government compiles. It is necessary for a combat aircraft in the military to be able to analyse petabytes of data.

Big data is used by government agencies to manage several agencies, such as dealing with traffic bottlenecks, controlling utilities and addressing the effects of crimes such as hacking and online fraud.

Aadhar Card: According to the government's records, there are 1.21 billion people. In order to determine things like the total number of young people in the country, this massive amount of data is examined and stored. Some plans are constructed to target the greatest number of people possible. As big data cannot be stored in a conventional database, it is necessary to employ the Big Data Analytics technologies in order to store and analyse the data.

E-commerce

E One of the applications of big data is also online shopping. It does things to preserve ties with consumers, which are extremely important for the e-commerce business. E-commerce websites have many different marketing ideas to retail merchandise clients, as well as ideas for managing transactions and implementing better tactics of new ideas to improve businesses using big data.

o Amazon: Amazon is a massive online retailer that processes a great quantity of daily traffic on its website. Yet, when Amazon has a sale that has been publicised in advance, there is a significant surge in traffic, which might cause the website to become unresponsive. Thus, it utilises Big Data in order to manage this kind of traffic and data. Big Data is helpful in arranging and evaluating the data for its future applications.

Social Media



The most significant contributor of data is social media. According to the statistics, social media platforms like Facebook create over 500 terabytes of new data every single day. This is especially true for Facebook. The majority of the data consists of films, images and other communication exchanges, among other things. A single action carried out on the social media website creates many data, which are then saved and processed only when necessary. The amount of data that is kept is measured in terabytes (TB) and processing it takes a significant amount of time. The solution to the issue is found in "Big Data."

5.2.4 Designing Data Architecture, R Syntax

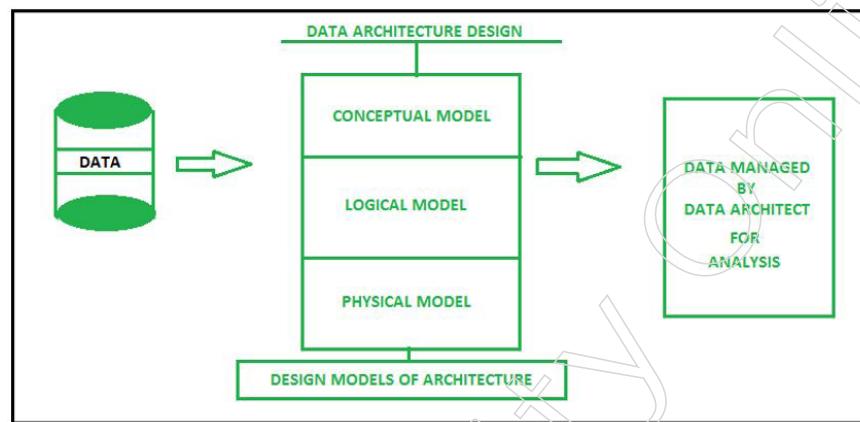
- **Data architecture Design**

Data architecture design is a set of standards that is composed of certain policies, rules, models and standards that manages what kind of data is collected, from where it is collected, the arrangement of collected data, storing that data, utilising and securing the data into the systems and data warehouses for further analysis. These standards can be broken down into four categories: policies, rules, models and standards. Data is one of the important foundations of enterprise architecture and it is one of the main reasons why the firm is successful in carrying out its business plan.

If a data architect wants to implement data integration, for instance, then it will require interaction between two systems and by using data architecture, the visionary model of data interaction during the process can be achieved. Data architecture design is important for creating a vision of interactions that occur between data systems. For example, if a data architect wants to implement data integration, then it will need interaction between two systems.

Notes

Data architecture not only simplifies the process of data preparation but also specifies the many types of data structures that may be utilised in the management of data. The data architecture is created by first splitting the data into three fundamental models and then combining those models together:



- **Conceptual model** – This is a type of business model that makes use of the Entity Relationship (ER) model to determine the relation that exists between entities and the attributes of those entities.
- **Logical model** - This type of model is one in which issues are depicted in the form of logic, such as rows and columns of data, classes, xml tags and other DBMS strategies.
- **Physical model** - The physical model stores the database design, such as which kind of database technology would be appropriate for the architecture. Physical models are also known as “physical representations.”

A data architect is the one who is accountable for the whole design, development, management and deployment of data architecture. This individual also determines how data is to be saved and accessed, while other decisions are made by internal bodies.

Factors that influence Data Architecture

There are a number of factors that might have an impact on data architecture, including corporate regulations, business requirements, the technology that is utilised, the economy and the demands of data processing.

- **Business requirements** – These include aspects such as the growth of the company, the efficiency with which users can access the system, data management, transaction management and the utilisation of raw data through the transformation of said data into image files and records, which is followed by their storage in data warehouses. The primary part of storing commercial transactions is done through the usage of data warehouses.
- **Business policies** – These policies, which are essentially a set of rules, are helpful for explaining the manner in which data is processed. Internal organisational entities in addition to other government agencies are responsible for formulating these policies.
- **Technology in use** – This may be demonstrated by using an example of a previously finished data architecture design, in addition to making use of any licenced software purchases or database technology already in place.

- **Business economics** – The design architecture will also be impacted by economic issues such company growth and loss, interest rates, loans, the condition of the market and the overall cost of the project.
- **Data processing needs** – The requirements for processing the data include aspects such as the mining of the data, massive continuous transactions, database administration and the requirements for any other necessary data preparation.
- **R Syntax**

The R programming language is becoming increasingly popular and is being utilised in a variety of data analysis applications. The manner in which we define its code is one that is not overly complicated. The “Hello World!” programme serves as the foundation for all programming languages; in this lesson, we will learn the syntax of the R programming language by utilising the “Hello World!” programme. Either the command prompt or a R script file might be utilised for the coding process that we undertake.

Syntax of R program

Variables, Comments and Keywords are the three components that make up a R programme. Comments are used to enhance the readability of the code, while keywords are reserved words that have a specific meaning to the compiler. Variables are used to store the data. Comments are used to improve the readability of the code.

Variables in R

In the past, we wrote all of our code inside of a print() function, but we do not currently have a method to address them in order to carry out further activities. This issue may be resolved by making use of variables, which, just like in any other programming language, are the names given to memory places that are designated specifically for the purpose of storing data of any kind.

There are three ways in which the assignment can be denoted in R:

1. = (Simple Assignment)
2. <- (Leftward Assignment)
3. -> (Rightward Assignment)

Example:



```
myFile.R*
1 var1 = "Simple Assignment"
2
3 var2 <- "Leftward Assignment!"
4
5 "Rightward Assignment" -> var3
6
7 print(var1)
8 print(var2)
9 print(var3)
10
```

Output:

“Simple Assignment”

Notes

“Leftward Assignment!”

“Rightward Assignment”

Comments in R

Your code's readability can be improved by the inclusion of comments, which are exclusively intended for the user and are thus ignored by the interpreter. Although R only supports single-line comments, it is possible to utilise multiline comments by employing a straightforward workaround, which will be explained in more detail below. Comments on a single line can be written by inserting a hash symbol (#) at the beginning of the sentence.

Example:



```
myFile.R* 
Source on Save Run Source
1 # This is a single line comment
2 print("This is fun!")
3
4 if(FALSE)
5 {
6   "This is multi-line comment which should be put inside either a
7     single or a double quote"
8 }
9
```

Output:

```
[1] "This is fun!"
```

From the above output, we can see that both comments were ignored by the interpreter.

Keywords in R

As a result of the unique significance attached to them, a programme will not allow a keyword to be used anywhere else in the code, even as the name of a variable or a function, for example.

We can view these keywords by using either `help(reserved)` or `?reserved`.

Reserved words in R

if	else	while	repeat	for
function	in	next	break	TRUE
FALSE	NULL	Inf	NaN	NA
NA_integer_	NA_real_	NA_complex_	NA_character_	...

- Control-flow statements and user-defined functions can be declared with the keywords if, else, repeat, while, function, for, in, next and break. Other control-flow statements include repeat and while.
- The ones that are still around are the ones that are utilised as constants, such as how TRUE and FALSE are used as boolean constants.
- The value Not a Number is specified by the NaN notation and the NULL notation is used to describe an undefined value.

- The value "inf" denotes an infinite amount.
- Note: R is a case sensitive language so TRUE is not same as True.

5.2.5 IDE for Hadoop, Integration with Big Data, Integration Methods

• Integration with Big Data

Integration of data is now standard procedure in every type of business. It is essential that data be safeguarded, managed, transformed, made useable and adaptable. Data underpins not just everything that we do on a personal level but also the capacity of businesses and other institutions to provide us with goods and services.

Big data integration refers to the process of utilising people, processes, suppliers and technology in a coordinated manner in order to gather, reconcile and make more effective use of data coming from a variety of sources in order to assist decision making. Big data may be characterised by its volume, velocity, truthfulness, variability, value and visibility. In addition, big data can be quite valuable.

- **Volume** – It is the primary characteristic that distinguishes big data from typical structured data that is organised and maintained by relational database management systems. When compared to the traditional method of handling data inputs, the number of data sources available is significantly greater.
- **Velocity** - An increase in the pace of data creation caused by the data source. The creation of data comes from a great number of sources and it can take on a variety of forms and unformatted structures.
- **Veracity** - Reliability of data, not all data has value, data quality concerns.
- **Variability** - The data must be managed from a variety of sources due to the fact that they are inconsistent.
- **Value** - Data must have value for processing; all data does not have value.
- **Visualisation** - Information must have significance and be easily comprehended by the user.

The integration of large amounts of data should be used to support all of your organisation's services. Your company should function as a high-performing team that exchanges data, information and expertise to facilitate the purchase decisions that your clients make regarding the services and products you offer.

Big Data Integration Process

Integration and processing of large amounts of data are essential for all of the data that is acquired. In order to support the end result that will be achieved through the employment of the data, the data must have value. Many businesses rely on big data scientists, analysts and engineers to help create value from the data they receive and analyse since so much data is collected from so many different sources. These professionals employ algorithms and other ways to assist in this endeavour.

The processing of big data needs to be compatible with the corporate governance rules in order to be successful. Ensure that the risk associated with making judgements based on the data is reduced. Contribute to the development and empowerment of

Notes

the organisation. Cut costs or keep them from rising. Enhance the effectiveness of operations as well as decision support.

The fundamental steps are as follows:

- Extract data from multiple sources.
- Store data in a suitable form.
- Analyse the data while transforming it and integrating it.
- Coordinate the loading and use of data.

Automating the process of data orchestration as well as data loading into apps is absolutely necessary for success. The organisation's capacity to effectively use big data will be hindered by the adoption of technology that does not provide simplicity of use, which will make the technology onerous.

• **Integration Methods**

The act of merging data from many sources is known as data integration. This helps data managers and executives assess the combined data and come to more informed judgements about their businesses. Finding the data, retrieving it, cleaning it up and presenting it are all steps in this process, which may be performed by a human or a computer system.

In order to get insights into business information, data managers and/or analysts might execute queries against the integrated data. Because there are so many possible benefits, organisations need to take the time to ensure that their objectives are aligned with the appropriate strategy. Let us go over the five different kinds of data integration so you can gain a better grasp on the topic (sometimes referred to as approaches or techniques). We will talk about the benefits and drawbacks of each type, as well as when to utilise each one.

1. **Manual data integration**

In manual data integration, a data manager is responsible for supervising all stages of the integration process, which typically involves creating custom code. Without the use of automation, this involves connecting the many sources of data, collecting the data and cleaning it, among other tasks.

The following are some of the advantages:

- **Reduced Cost:** This method requires very little maintenance and, in most cases, can only combine a limited number of data sources. As a result, it has a lower overall cost.
- **Greater Freedom:** The individual using the application retains complete control over the integration.

Some of the cons are:

Some of the negatives are as follows:

- **Less Access:** There is less access available since a developer or management needs to manually coordinate each integration.

- **Difficulty Scaling:** Scaling is difficult because it requires manually updating the code for each integration, which is a time-consuming process. This makes scaling difficult for bigger projects.
- **Greater Room for Error:** There is a greater potential for mistake since the data must be handled by a management and/or an analyst at each level.

As it is a highly laborious and manual procedure, this method is most effective when applied to one-time occasions. But, it soon loses its viability when applied to complicated or ongoing integrations. Everything, from the collecting of data to its cleansing to its display, is done manually, which means that the processes involved require time and money.

2. Middleware data integration

The term “middleware” refers to software that is used to connect applications and facilitate the transmission of data between such apps and databases. Since it may serve as a translator between different computer systems, middleware is particularly useful for companies that are attempting to combine recalcitrant old systems with more modern ones.

The following are some of the advantages:

- **Better data streaming:** Improved data flow as a result of the software’s ability to carry out integration in a manner that is both automated and consistent every time.
- **Easier access between systems:** Since the software is intended to make communication easier across the systems in a network, users will find that accessing other systems is much simpler.

Some of the cons are:

- **Less Access:** There is less access available since the middleware must be deployed and maintained by a developer who has some level of technical expertise.
- **Limited Functionality:** Middleware has limited capability since it can only interact with particular types of computer systems.

Middleware is suitable for companies who are connecting ancient systems with more current systems. Nevertheless, middleware is mostly used as a communications tool and has limited capabilities for data analytics.

3. Application-based integration

With this method, the task is carried out entirely by various software programmes. They are responsible for locating, retrieving, cleaning and integrating data from a variety of sources. Because of this interoperability, it is quite simple for data to be transferred from one source to another.

The following are some of the advantages:

- **Simplified procedures:** All of the work may be completed automatically by a single application.
- **Facilitated information exchange:** The application enables information to be transferred across systems and departments in a smooth manner.

Notes

- **Less consumption of available resources:** The fact that a significant portion of the process is automated frees up managers and analysts to work on other tasks.

The following are some of the disadvantages:

- **Restricted access:** This method necessitates specialised, technical expertise in addition to the employment of a data manager and/or analyst to supervise the implementation and upkeep of the programme.
- **Inconsistent results:** The method is not standardised and differs considerably amongst companies who provide this service to customers.
- **Complex setup:** In order to design the application(s) so that they function seamlessly across departments, it is necessary to have developers, managers and/or analysts who are knowledgeable in technical matters.
- **Difficulty in managing data:** Having access to several systems might result in the integrity of the data being compromised.

Due to its prevalence among businesses that operate in hybrid cloud environments, this strategy is sometimes referred to as enterprise application integration. These types of companies are required to collaborate with a variety of data sources, both on-premises and in the cloud. This strategy improves the efficiency of data transfer and processes between the two environments.

4. Uniform access integration

With this method, data may be accessed from even more distinct collections and then it is presented in a standard format. This is accomplished while the data are allowed to remain in their previously established position.

The following are some of the benefits:

- **Reduced storage needs:** There is no need to build a separate location to store data, therefore this eliminates one of the storage requirements.
- **Easy access to the data:** This method works well with a variety of different systems and sources of data.
- **Simplified view of the data:** This method gives the end user a consistent representation of the data by simplifying its look.

The following are some of the challenges:

- **Data Integrity Challenges:** Accessing such a large number of sources might result in the integrity of the data being compromised.
- **Strained Systems:** Data host systems are not often built to be able to deal with the volume and frequency of data requests that occur throughout this process.

This strategy is the best option for companies that need access to a variety of different computer systems. This strategy has the potential to produce insights without the expense of generating a backup or duplicate of the data, provided that the data request does not place an undue stress on the host system.

5. Common storage integration (sometimes referred to as data warehousing)

This strategy is quite similar to the uniform access method, with the exception that it requires a copy of the data to be created and kept in a data warehouse. Because

of this, organisations are able to alter data in a more versatile manner, which has contributed to its status as one of the most popular kinds of data integration.

The following are some of the advantages:

- **Reduced burden:** There is not a continuous cycle of data requests being processed by the host system.
- **Increased data version management control:** A higher level of data integrity may be achieved by accessing the data from a single source rather than several independent sources.
- **Cleaner data appearance:** It is possible for managers and/or analysts to conduct a variety of queries on the copy of the data that has been saved while still keeping the data's presentation consistent.
- **Enhanced data analytics:** The management and/or analysts are able to conduct more complex queries thanks to the existence of a stored copy, which eliminates any concerns about the data's integrity being jeopardised.

Some of the cons include:

- **Increased storage costs:** If you want to create a duplicate of the data, you will need to locate and pay for a location to keep it.
- **Higher maintenance costs:** In order to orchestrate this strategy, technical professionals are required to set up the integration, supervise it and keep it maintained.

The most cutting-edge method of integration is the utilisation of shared storage. Because it enables the most complex querying, this strategy is almost probably the most effective option for companies to pursue if they have the capacity to do so. This level of expertise can lead to a greater depth of understanding.

5.3 Introduction to Neural Networks

Introduction

A neural network is a collection of neurons that, in combination with information from other nodes and based on the input they receive, generate output without following any predetermined rules. In essence, they approach the resolution of issues by a process of trial and error. The human and animal brains serve as inspiration for neural networks. Although if neural networks are already sophisticated enough to beat human opponents in games like as chess and GoExternal link:open in new, they still do not have the same level of cognitive ability as a human child or the majority of other animals.

5.3.1 Introduction to Neural Network

In the field of artificial intelligence, a neural network is a technique that instructs computers to interpret data in a manner that is modelled after how the human brain does so. Deep learning is a form of machine learning method that emulates the layered structure of the human brain via the use of linked nodes or neurons. This approach is also known as neural network learning. It does this by establishing an adaptable framework that allows computers to gain knowledge from their past errors and

Notes

constantly improve. Hence, artificial neural networks are used in an effort to handle complex issues, such as summarising papers or identifying faces, with a better degree of precision.

The use of neural networks enables computers to make intelligent judgements with relatively little input from humans. This is due to the fact that they are able to learn and understand the complicated nonlinear interactions that exist between the input data and the output data. For example, they are capable of doing the activities listed below:

Make generalizations and inferences

Unstructured data may be comprehended by neural networks and they can make broad observations even without receiving any specific training. For example, they are able to distinguish that two separate sentences in the input have a meaning that is comparable to one another:

- Would you be able to instruct me on how to make the payment?
- What are the steps to transferring money?

A neural network would recognise that both phrases imply the same thing since they are semantically equivalent. Alternatively, it would be able to understand in a general sense that although Baxter Road is a location, Baxter Smith is a name of a person.

There are a variety of applications for neural networks across a wide range of sectors, including the following:

- Medical diagnosis by the categorization of medical images.
- Targeted marketing through the filtering of social networks and the study of behavioural data.
- Predictions of the economy based on analysis of past data relating to financial instruments.
- Electricity load and energy demand forecasting.
- Control of the manufacturing process and quality.
- Identification of chemical compounds.

Here you will find a discussion of four of the most significant uses of neural networks.

Computer vision

The capacity of computers to glean information and insights from still photos and moving films is referred to as "computer vision." Neural networks enable computers to differentiate and recognise pictures in a manner analogous to that of humans. There are many different uses for computer vision, including the following:

- Visual recognition technology in driverless automobiles, so that they can distinguish road signs and other vehicles on the road.
- Content moderation, which may automatically delete any content from picture and video archives that is deemed to be harmful or improper.
- Recognition of faces using facial features, such as open eyes, spectacles and facial hair, to facilitate identification and recognition.

- Image labelling for the purpose of identifying company logos, clothes, safety gear and other image details.

Speech recognition

Although though people have diverse speech patterns, pitches, tones, languages and accents, neural networks are able to interpret human speech. Speech recognition is used to do a variety of activities, such as those listed below, by virtual assistants such as Amazon Alexa and automatic transcribing software.

- Support employees working in call centres and automatically sort incoming calls.
- Convert clinical interactions into documentation in real time.
- Subtitle films and recordings of meetings in an accurate manner so that more people may access the information.

Natural language processing

The capacity to process text that was written naturally by humans is known as natural language processing, or NLP. Computers are able to glean insights and meaning from text data and documents with the assistance of neural networks. There are many applications for natural language processing, such as the following functions:

- Computer-generated chatbots and automated virtual agents.
- The organising and categorization of written data on an automated basis.
- A study of long-form documents such as emails and forms performed by a business intelligence team.
- The indexing of terms that are important in determining emotion, such as favourable and negative comments made on social media.
- Synopsis of relevant documents and creation of articles based on a specified subject.

Recommendation engines

User behaviour may be tracked by neural networks, which can then be used to produce customised suggestions. They are also able to monitor the activity of all users and find new products and services that a particular user might be interested in. For instance, Curalate, a firm located in Philadelphia, assists companies in turning the engagement they receive on social media into actual revenue. The intelligent product tagging (IPT) solution offered by Curalate is utilised by companies in order to automate the collecting and curation of user-generated social material. IPT makes use of neural networks to automatically locate and propose things to the user that are related to the user's behaviour on social media. Customers no longer have to sift through many web catalogues in order to identify a product that they saw advertised on social media. They might, alternatively, make advantage of Curalate's automatic product labelling in order to make the purchase of the product more easily.

What are the types of neural networks?

The manner in which data moves from the input node to the output node may be used to classify different types of artificial neural networks. Several instances are listed below:

Notes

Feedforward neural networks

The data is processed in feedforward neural networks in just one way and that is from the input node to the output node. Every single node on one layer is linked to each and every node on the layer above it. The accuracy of predictions made by a feedforward network may be improved with the help of a feedback mechanism.

Backpropagation algorithm

The predicted accuracy of artificial neural networks may be continually improved by employing corrective feedback loops in their learning processes. You may conceive of the data as travelling from the input node to the output node in the neural network over a variety of different pathways. Here is a simplified explanation of how the process works. There is just one path that should be taken in order to correctly translate the input node to the output node. In order for the neural network to locate this route, it makes use of a feedback loop, which operates as follows:

1. Each node in the path makes an educated judgement as to which node will come next in the path.
2. It determines whether or not the prediction was accurate. Paths that result in fewer inaccurate predictions have lower weight values assigned by the nodes, whereas pathways that result in more right guesses receive higher weight values assigned by the nodes.
3. The nodes will then generate a new forecast for the next data point by utilising the pathways with the higher weights and then Step 1 will be repeated.

Convolutional neural networks

Convolutions are the name for the unique mathematical operations that are carried out by the hidden layers of convolutional neural networks. These operations include filtering and summarising data. They can extract significant characteristics from images that are useful for image recognition and classification, which makes them particularly valuable for image classification. The new form can be processed more quickly without sacrificing any of the important aspects that are necessary for producing an accurate forecast. A variety of picture characteristics, including edges, colour and depth, are extracted and processed by each of the hidden layers.

5.3.2 Difference Between Human Brain and Artificial Network

1. **Artificial Neural Network:** An artificial neural network, often known as an ANN, is a sort of neural network that uses a Feed-Forward approach to process information. It is referred to as such because the information is continually sent across all of the nodes up until it reaches the output node. One might also refer to this as the most basic variety of neural network.

ANN has a number of benefits, including the following:

- Capability to learn regardless of the type of data (Linear or Non-Linear).
- ANN is extremely volatile and its use shines brightest in the analysis of financial time series.

There are a few drawbacks associated with the ANN, including the following:

- Since it has the simplest architecture, it is difficult to describe the behaviour of the network.
- This network is dependent on hardware.

2. Biological Neural Network: The Biological Neural Network (BNN) is a structure that is made up of the synapse, dendrites, cell body and axon of a neuron. Neurons, which are part of this neural network, are the ones doing the processing. Dendrites are responsible for receiving signals from other neurons, the soma region is responsible for adding up all of the incoming signals and axons are responsible for transmitting the signals to other cells.

BNN has a number of benefits, including the following:

- The synapses are the element responsible for input processing.
- It is able to process very complicated parallel inputs.

Among the many drawbacks of BNN are the following:

- There is no regulating mechanism.
- The speed of processing is slow due to the complexity of the situation.

Differences between ANN and BNN

There are some distinctions between Biological Neural Networks (BNNs) and Artificial Neural Networks (ANNs), despite the fact that both types of networks are made up of fundamental components that are quite similar to one another.

- **Neurons:** In both BNNs and ANNs, neurons are the fundamental components that are responsible for information processing and transmission. Yet, the neurons of a BNN are more complicated and varied than those of an ANN. With BNNs, neurons take input from various sources via their multiple dendrites, while the axons convey signals to other neurons. In ANNs, on the other hand, neurons are simplified and often only have a single output.
- **Synapses:** Synapses are the places of connection between neurons in both BNNs and ANNs. It is at these synapses that information is conveyed. On the other hand, in ANNs, the connections between neurons are typically fixed and the strength of the connections is determined by a set of weights. On the other hand, in BNNs, the connections between neurons are more flexible and the strength of the connections can be modified by a variety of factors, including learning and experience.
- **Neural Pathways:** Neural routes are the connections between neurons that are present in both BNNs and ANNs. These connections are what enable information to be transported throughout the network. On the other hand, the neuronal pathways in BNNs are extremely intricate and varied and the connections between neurons are malleable thanks to the effects of experience and learning. In artificial neural networks, the neural pathways are often more straightforward and are predetermined by the network's architecture.

Size: the human brain is estimated to have around 86 billion neurons and more than 100 trillion synapses (or, according to some calculations, 1000

Notes

trillion synapses) (connections). It is incorrect to compare their numbers in this way because the number of “neurons” in artificial networks is far lower (often in the range of 10–1000), yet the comparison is nevertheless made. The only thing that happens in perceptrons is that their “dendrites” accept input and their “axon branches” create output. A single layer perceptron network is comprised of several perceptrons, but these perceptrons are not coupled to one another in any way; rather, they all carry out the same activity simultaneously. Deep Neural Networks typically have input neurons, which can be as numerous as the number of features in the data, output neurons, which can be as numerous as the number of classes if the network was built to solve a classification problem and neurons in the hidden layers, which are located in-between the other two types of neurons. It is typical for each layer to be fully linked to the layer below it, however this is not always the case. This means that artificial neurons typically have the same number of connections as there are artificial neurons in both the layer below them and the layer above them combined. Convolutional Neural Networks are able to extract characteristics from the data using a variety of methods, each of which is more complex than what can be accomplished by a small number of linked neurons working alone. Manual feature extraction, which involves modifying data in such a manner that it can be fed to machine learning algorithms, requires the mental capacity of a human, which is another factor that is not taken into consideration when calculating the total number of “neurons” necessary for Deep Learning tasks. The constraint in size is not only a technical one: merely increasing the number of layers and artificial neurons does not necessarily result in improved outcomes when applied to machine learning applications.

- **Topology:** The topology is such that all artificial layers do computations in order, rather than being a component of a network that consists of nodes that perform computations in an asynchronous manner. The state of one layer of artificial neurons and the weights of those neurons are computed using feedforward networks, which then use the results to calculate the next layer in the same manner as before. During backpropagation, the algorithm computes some change in the weights that go in the opposite direction, with the goal of minimising the gap between the feedforward computational results in the output layer and the expected values of the output layer. This is accomplished by changing the weights in the opposite direction. Although layers are not connected to one other or to layers that aren’t immediately adjacent, it is feasible to simulate loops with recurrent and LSTM networks. Neurons in biological networks can fire asynchronously in parallel and have a small-world structure, with a small percentage of highly linked neurons (hubs) and a large percentage of neurons with less connections to one another (the degree distribution at least partly follows the power-law). This small-world characteristic of biological neurons can only be mimicked by inserting weights that are 0, which mimics the lack of connections between two neurons. This is because artificial neuron layers are often completely linked, therefore this is the only way to emulate it.

5.3.3 Perceptron Model: Its Features, McCulloch Pitts Model

- **Perceptron Model**

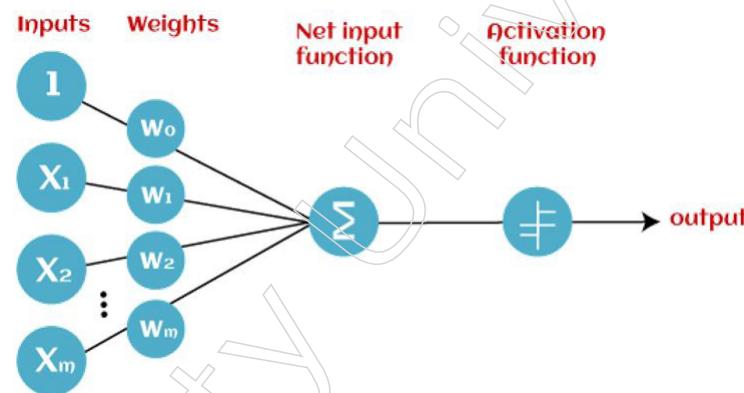
In addition, the term “Perceptron” can also refer to a “Artificial Neuron” or a “neural network unit,” both of which are utilised in business intelligence to assist in the detection of particular input data calculations. The Perceptron model is often considered to be among the most effective and user-friendly varieties of artificial neural networks. Nonetheless, it is a binary classifier learning algorithm that uses supervised learning. As a result, we may think of it as a neural network with a single layer and four primary parameters, namely input values, weights and bias, net sum and an activation function.

What is Binary classifier in Machine Learning?

In the field of machine learning, binary classifiers are functions that assist in determining whether or not the input data can be represented as vectors of numbers and whether or not it belongs to a certain category. It is possible to think of binary classifiers in terms of linear classifiers. We may consider it as a classification method that can predict linear predictor function in terms of weight and feature vectors. To put it another way, we can think of it as simple words.

Basic Components of Perceptron

Mr. Frank Rosenblatt was the one who came up with the idea for the perceptron model, which is a binary classifier and has three primary components. The following describe each of these:



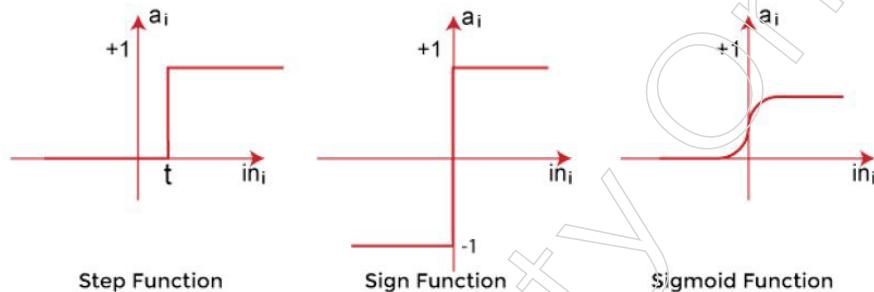
- **Input Nodes or Input Layer:** This is the fundamental component of the Perceptron system that is responsible for accepting the initial data into the system so that it may be processed further. Each input node carries a true numerical value.
- **Weight and Bias:** The Weight parameter describes the strength of the link between units. Bias represents how strongly one unit is connected to another. Another of the Perceptron components' most crucial parameters is the value of this variable. When it comes to determining the output, the intensity of the linked input neuron is precisely proportional to the weight of the object. In addition, one may think of the bias as the line that intercepts the variable in a linear equation.
- **Activation Function:** These are the last and most crucial components that aid to determine whether the neuron will fire or not. They are responsible for

Notes

determining whether or not the neuron will fire. A step function is the most appropriate way to conceptualise the Activation Function.

Types of Activation functions:

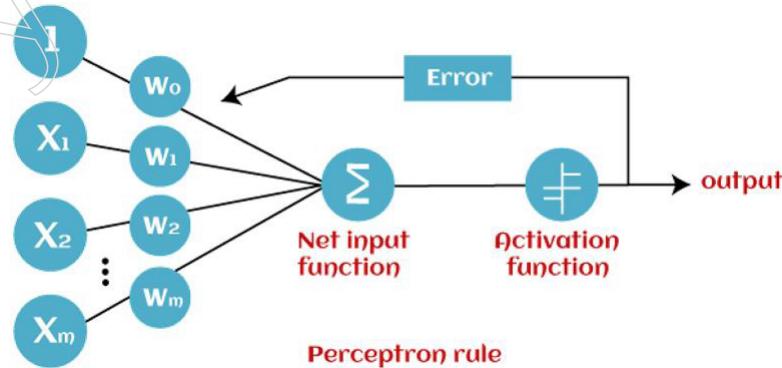
- Sign function
- Step function and
- Sigmoid function



The data scientist makes the desired outputs by using the activation function to arrive at a conclusion based on their own personal preferences, which is then informed by the numerous issue statements. In perceptron models, the activation function may be different (for example, Sign, Step, or Sigmoid) depending on whether or not the learning process is sluggish, or whether or not it has vanishing or exploding gradients.

How does Perceptron work?

When it comes to machine learning, a perceptron is referred to as a single-layer neural network. This type of neural network has four primary parameters, which are referred to as input values (Input nodes), weights and Bias, net sum and an activation function. The perceptron model starts by multiplying all of the input values by their respective weights, then it adds all of these values together to form the weighted total of all of the input values. The required output may then be obtained by applying this weighted sum to the activation function denoted by the letter f . This activation function, which may also be referred to as the step function, is denoted by the letter ' f ' in mathematical notation.



This step function or Activation function plays a vital role in ensuring that output is mapped between required values (0,1) or (-1,1). It is essential to keep in mind that the quantity of input serves as an indication of the power of a node. In a similar manner, the value of the bias input provides the opportunity to move the activation function curve either upwards or downwards.

The following are the two significant steps involved in the operation of the perceptron model:

Step-1

In the first step, first multiply all of the input values by their respective weight values, then add up all of the products to get the weighted total. The following is the mathematical formula that may be used to determine the weighted sum:

$$\sum w_i * x_i = x_1 * w_1 + x_2 * w_2 + \dots + x_n * w_n$$

To enhance the performance of the model, a specialised term known as bias 'b' should be included to the weighted sum.

$$\sum w_i * x_i + b$$

Step-2

In the second stage, an activation function is applied together with the weighted sum that was discussed before. This provides us with output that is either in the form of binary digits or a continuous number as described below:

$$Y = f(\sum w_i * x_i + b)$$

Types of Perceptron Models

There are two different kinds of Perceptron models and you can tell them apart by the layers. The following describe each of these:

1. Single-layer Perceptron Model
2. Multi-layer Perceptron model

Single Layer Perceptron Model:

This particular form of artificial neural network (ANN) is one of the more straightforward ones. A feed-forward network and a threshold transfer function are both components of a single-layered perceptron model. The model also incorporates both of these components. The analysis of linearly separable objects with binary outcomes is the primary purpose of the single-layer perceptron model.

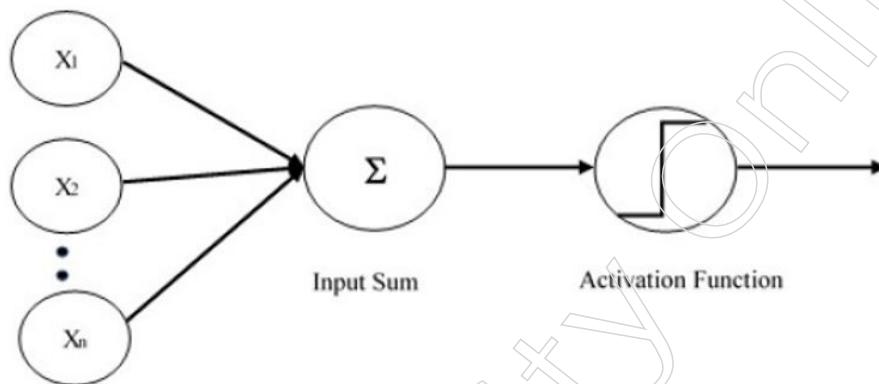
Since the algorithms that make up a model with a single layer perceptron do not contain any recorded data, the model starts with input that is inconsistently assigned for the weight parameters. In addition, it compiles all of the inputs (weight). After combining all of the inputs together, the model is activated and displays the output result as +1 if the entire sum of all inputs is greater than a value that has been specified in advance.

If the outcome is the same as the value that was pre-determined or the threshold value, then the performance of this model is said to be fulfilled and there will be no change in the need for weight. Nevertheless, this model has a few inconsistencies that become apparent when it is fed various values for the weight inputs. These inconsistencies are triggered when the model is run with multiple weight inputs. As a result, in order to get the intended output and reduce the number of mistakes, the weights input can require some adjustments.

Notes

Notes

Single layer perceptron was the first neural model to be proposed. The content of the neuron's local memory consists of a vector of weights. The computation of a single-layer perceptron involves the calculation of the sum of the input vector multiplied by the corresponding element of the weights vector. The value displayed on the output will be the input to an activation function.



"Single-layer perceptron are only capable of learning patterns that can be separated linearly."

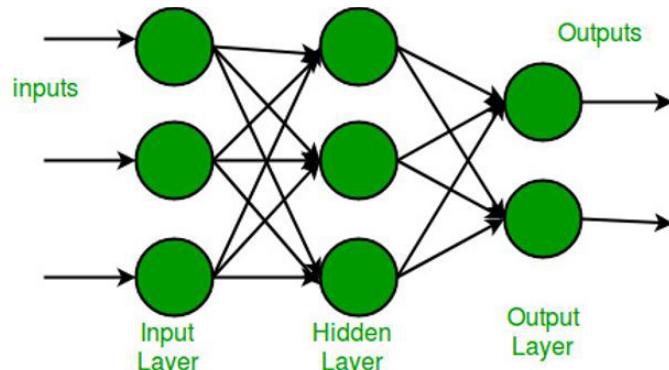
Multi-Layered Perceptron Model:

A multi-layered perceptron model is quite similar to a single-layer perceptron model; however, a multi-layered perceptron model contains a higher number of hidden layers than a single-layer perceptron model does. In addition to its other name, the multi-layer perceptron model is also known as the Backpropagation method. This algorithm operates in two phases and the stages are as follows:

- **Forward Stage:** The activation functions for this stage begin on the input layer and end on the output layer. This stage is also known as the "forward" stage.
- **Backward Stage:** The backward stage is when the model's weight and bias values are adjusted in accordance with the requirements of the model. At this point, the difference between the actual production that was produced and the demand for that output began in reverse on the output layer and concluded on the input layer.

As a result, a multi-layered perceptron model has been viewed as many artificial neural networks consisting of various layers within which the activation function does not stay linear. This is analogous to a single layer perceptron model. For deployment purposes, the activation function can be run as a sigmoid, TanH, ReLU, or any other function except linear. A multi-layer perceptron model has increased processing capability and is capable of processing both linear and non-linear patterns. In addition to that, it has the capability of implementing logic gates such as AND, OR, XOR, NAND, NOT, XNOR and NOR.

A multi-layer perceptron has one input layer with one neuron (or node) for each input, one output layer with one neuron (or node) for each output, and any number of hidden layers and any number of nodes for each hidden layer. Below is a schematic representation of a Multi-Layer Perceptron (MLP).

**Notes**

In the above multi-layer perceptron diagram, we can see that there are three inputs and, consequently, three input nodes, as well as three hidden layer nodes. There are two output nodes because the output component provides two outputs. The nodes in the input layer accept input and forward it for further processing. In the diagram, the nodes in the input layer forward their output to each of the three nodes in the hidden layer, similarly, the hidden layer processes the information and passes it to the output layer.

Advantages of Multi-Layer Perceptron:

- It is possible to employ a multi-layered perceptron model to find solutions to complicated non-linear issues.
- It performs admirably with both limited and extensive amounts of input data.
- We are able to get accurate forecasts with its support following the training.
- This facilitates achieving the same accuracy ratio with huge data sets as well as with small data sets.

Disadvantages of Multi-Layer Perceptron:

- The computations involved in multi-layer perceptron are both challenging and time-consuming.
- It is difficult to make accurate projections about the degree to which the dependent variable impacts each independent variable when using a multi-layer Perceptron.
- The successful operation of the model is contingent on the calibre of the instruction.

Perceptron Function

The output of the perceptron function “ $f(x)$ ” may be obtained by multiplying the input value ‘ x ’ with the weight coefficient ‘ w ’ that was previously learnt.

To put it in mathematical terms, it may be stated as follows:

$$f(x)=1; \text{ if } w.x+b>0$$

$$\text{otherwise, } f(x)=0$$

- ‘ w ’ represents real-valued weights vector
- ‘ b ’ represents the bias
- ‘ x ’ represents a vector of input x values.

Notes

Characteristics of Perceptron

The perceptron model has the following characteristics.

1. The Perceptron is a method for the supervised learning of binary classifiers that is used in machine learning.
2. The weight coefficient is automatically learnt when using the perceptron algorithm.
3. At the beginning, weights are multiplied with the input characteristics and then a choice is made on whether or not the neuron will fire.
4. The activation function uses a step rule to determine if the weight function is positive or negative and then checks to see if it is larger than zero.
5. The linear decision border is established, which enables the distinction to be made between the two classes that may be linearly separated: +1 and -1.
6. It is required to have an output signal if the cumulative total of all of the input values is more than the threshold value; else, there will be no output shown.

Limitations of Perceptron Model

The following are some of the constraints of a perceptron model:

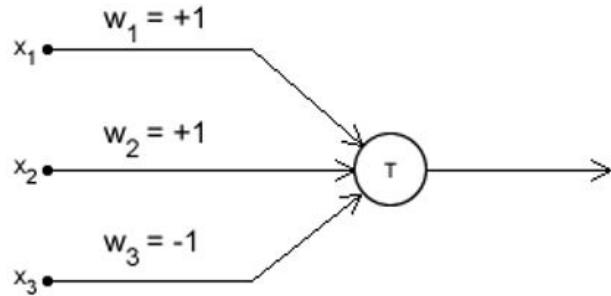
- Since the hard limit transfer function exists, the output of a perceptron can only ever be a binary integer (0 or 1).
 - The Perceptron algorithm can only be used to categorise sets of input vectors that can be separated linearly. When the input vectors are non-linear, it is difficult to appropriately categorise them because of their shape.
- **McCulloch Pitts Model**

An artificial neuron or perceptron, as it is more often known, is the most fundamental unit of deep neural networks. This fact is quite well known. However, McCulloch and Pitts made the very first step towards the perceptron that we use today in 1943 by imitating the functionality of a biological neuron. This was the initial step towards the perceptron that we use today.

The McCulloch-Pitts model depicted an artificial neuron that was quite straightforward in its construction. A zero or a one may have been used as one of the inputs. In addition, the output was either a 0 or a 1. And each input may either be excitatory or inhibitory, depending on the context.

So the objective at hand was to total up all of the inputs. When an input has a value of one and has an excitatory effect, we say that it added one. If it was one and it had an inhibitory effect, then it took one away from the total. After carrying out this process for each of the inputs, a total sum is computed. If this final total is less than a certain amount, which you choose (let us call it T), then the output will be zero. In all other cases, the outcome will be a 1.

The McCulloch-Pitts model is presented here in the form of a graphical depiction.



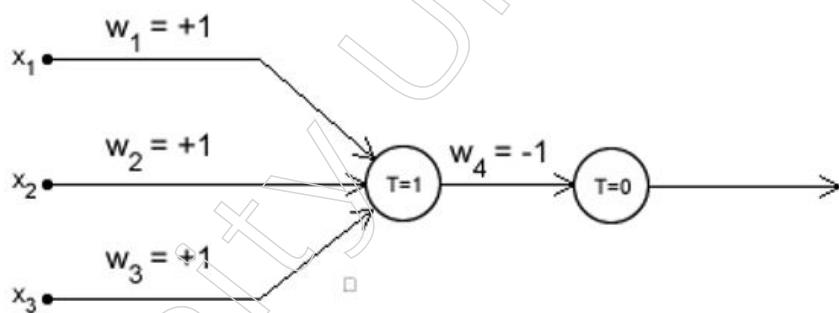
Various aspects of the graphic are represented using named variables. Excitatory input is indicated by the variables w_1 , w_2 and w_3 , whereas inhibitory input is shown by the third variable. These things are referred to as "weights." According to this concept, an excitatory input is a weight that has a value of 1. When it is -1, it indicates that the variable is an inhibitory input. The inputs are denoted by the variables x_1 , x_2 and x_3 . If it were necessary, there may be more (or fewer) inputs. And as a direct result of this, there would be an increased number of 'w's to indicate whether or not that specific input is excitatory or inhibitory. So, if you give it some thought, you'll see that you can compute the total by utilising the 'x's and the 'w's... something along these lines:

$$\text{sum} = x_1w_1 + x_2w_2 + x_3w_3 + \dots$$

This is what is called a 'weighted sum'.

Now that the total has been computed, we need to determine whether or not sum is less than T . If this is the case, the output will be set to zero. In such case, it will be counted as one. Now, by making use of this straightforward model of a neuron, we are able to get some fascinating results. The following are some examples:

NOR Gate

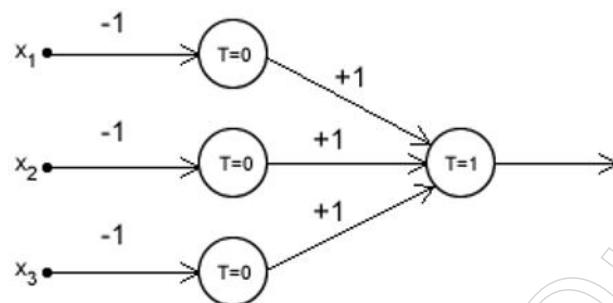


The illustration to the right depicts a NOR gate with three inputs. The only time a NOR gate will produce an output of one is if all of its inputs are 0. (in this case, x_1 , x_2 and x_3) You can experiment with the many potential combinations of inputs (they can be either zero or one).

Take note that there are two neurons being used in this illustration. The inputs that you provide are taken in by the first neurons. The output of the first neuron is used as the basis for the work of the second neuron. It has no idea what the primary inputs were to begin with.

Notes

NAND Gate



The following figure demonstrates how these neurons may be used to produce a NAND gate with three inputs. Only when all of its inputs are 1 does a NAND gate produce a zero. This neuron requires a total of four neurons. The output of the first three neurons is what the fourth neuron receives as its input. If you experiment with the many permutations of the inputs.

5.3.4 Role of Activation Function, Backpropagation Algorithm

- **Role of Activation Function**

In the process of building a neural network, one of the choices you get to make is what Activation Function to use in the hidden layer as well as at the output layer of the network.

Elements of a Neural Network

- **Input Layer:** This layer is responsible for receiving input features. No calculation is carried out at this layer; the nodes here just pass on the information (features) to the next layer, which is known as the hidden layer. It supplies the network with information from the outside world.
- **Hidden Layer:** Nodes of this layer are not exposed to the outer world, they are part of the abstraction provided by any neural network. The hidden layer performs all sorts of computation on the features entered through the input layer and transfers the result to the output layer.
- **Output Layer:** The information that has been learnt by the network is brought up to this layer, which is known as the output layer.

What is an activation function and why use them?

Calculating the weighted total and then adding bias to it allows the activation function to determine whether or not a neuron should be engaged. This decision is made by the activation function. The activation function of a neuron is designed to do one thing and that is to inject non-linearity into the output of the neuron.

Explanation: As is well knowledge, the neural network consists of neurons that function in correspondence with weight, bias and the activation function that is specific to each neuron. If we were using a neural network, we would adjust the weights and biases of the neurons based on the amount of error that was being produced by the network. Back-propagation is the name given to this particular mechanism. Back-propagation is only feasible because activation functions supply the gradients together with the error that has to be updated in order to bring the weights and biases up to date.

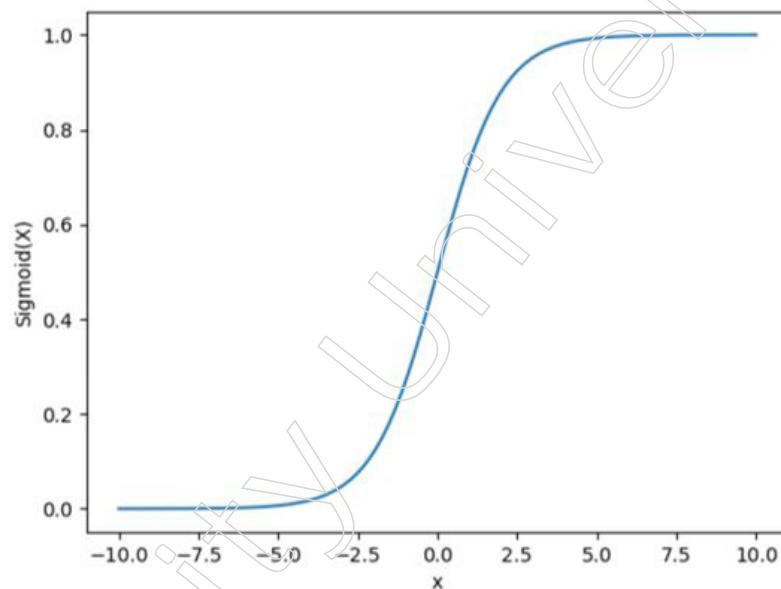
Variants of Activation Function

Linear Function

- **Equation:** Linear function has the equation similar to as of a straight line i.e. $y = x$
- No matter how many layers we have, if they are all linear in nature, the final activation function of the last layer is nothing but a linear function of the input of the first layer. This is true regardless of how many layers we have.
- **Range:** $-\infty$ to $+\infty$
- **Uses:** Linear activation function is used at just one place i.e. output layer.
- **Issues:** If we will differentiate linear function to bring non-linearity, result will no more depend on input "x" and function will become constant, it won't introduce any ground-breaking behaviour to our algorithm.

For example: Consider the calculation of the cost of a house; this is an example of a regression problem. Because the value of the house price might be either large or tiny, we can utilise linear activation for the output layer. Even in this particular scenario, the neural net has to have some kind of non-linear function at the hidden layers.

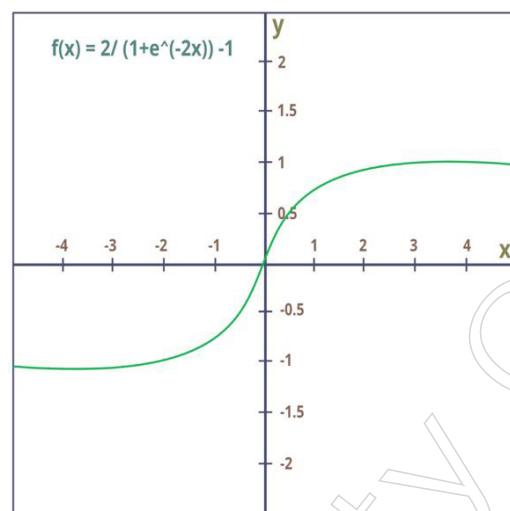
Sigmoid Function



- It is a function whose graph looks like a 'S' shape when it is plotted.
- **Equation :** $A = 1/(1 + e^{-x})$
- Its very nature is non-linear. Take note that the X values range from -2 to 2, while the Y values are quite steep. This indicates that even little alterations in the value of x might result in significant shifts in the value of Y.
- The range of values is from 0 to 1.
- **Applications:** Typically used in the output layer of a binary classification, where the result is either 0 or 1, as the value for a sigmoid function lies only between 0 and 1. Because of this, the result can be easily predicted to be 1 if the value is greater than 0.5 and it can be predicted to be 0 otherwise.

Notes

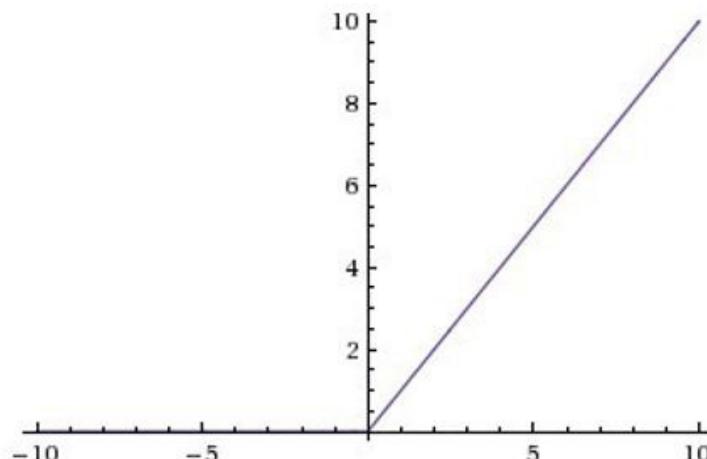
Tanh Function



DG

- The Tanh function, sometimes referred to as the Tangent Hyperbolic function, is the activation that nearly invariably performs better than the sigmoid function. In reality, it is a variant of the sigmoid function that has been mathematically shifted. Both are related to one another and may be constructed using the other.
 - **Equation :-**
- $$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$$
- **Value Range :-** -1 to +1
 - **Nature :-** non-linear
 - **Applications:** Often utilised in the hidden layers of a neural network due to the fact that its values range from -1 to 1, which causes the mean for the hidden layer to be either 0 or extremely near to reaching that value. This assists with data centring by bringing the mean closer to 0. The learning process for the subsequent layer is simplified as a result of this.

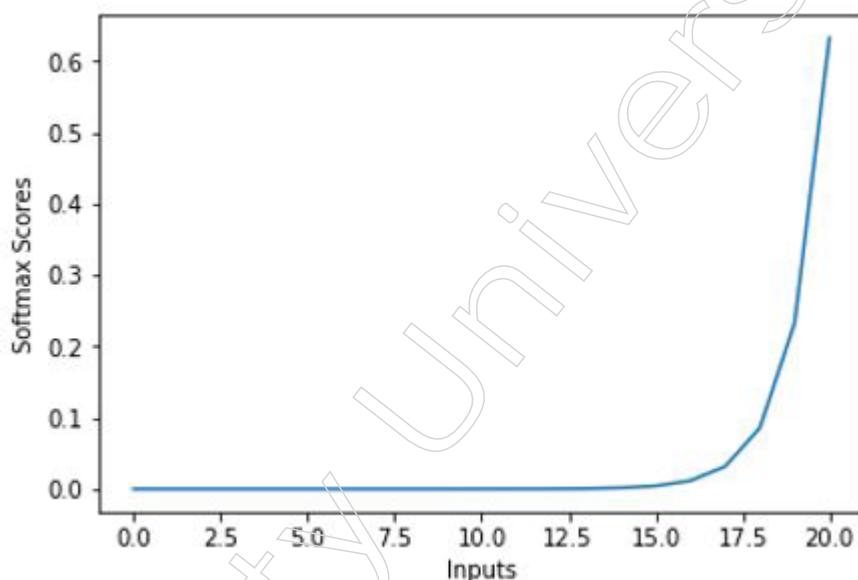
RELU Function



- Rectified linear unit is what it is short for. It is the activation function that is called upon the most frequently. The majority of the implementation takes place in the Neural network's hidden layers.
- **Equation:** $A(x) = \max(0, x)$. It gives an output x if x is positive and 0 otherwise.
- **Value Range:** $[0, \infty)$
- **Nature:** Non-linear in nature, which allows us to easily backpropagate mistakes and have several layers of neurons triggered by the ReLU function.
- The nature of the network is such that it is not linear.
- **Applications:** Since it relies on less complex mathematical operations, ReLU requires less processing power than alternative methods such as tanh and sigmoid. At any one time, only a few number of neurons are engaged, resulting in a sparse network that is both effective and simple to compute on.

To put it another way, RELU is capable of substantially quicker learning than sigmoid and Tanh functions.

Softmax Function



The softmax function is another kind of sigmoid function, however it comes in particularly helpful when we are attempting to deal with issues involving several classes of categorization.

- **Nature:** non-linear
- **Uses:** Often utilised while attempting to manage many courses at the same time. It was usual practise to use the softmax function in the output layer of picture classification issues. The softmax function would compress the outputs for each class to a range between 0 and 1 and it would also divide the total number of outputs by those outputs.
- **Output:** The softmax function is most effectively utilised in the output layer of the classifier, which is the part of the process in which we are attempting to arrive at the probabilities that will determine the category of each input.

Notes

- If the purpose of your output is binary classification, then the sigmoid function is an extremely obvious option for the output layer.
- If the outcome of your analysis is for multi-class classification, then the probabilities of each class may be predicted with great accuracy using Softmax.

• Backpropagation

A method known as backpropagation, also known as the backward propagation of mistakes, is meant to check for faults by moving in reverse order from the output nodes to the input nodes. In the fields of data mining and machine learning, it is an essential piece of mathematical software for enhancing the precision of forecasts. In its most basic form, backpropagation is an algorithm that facilitates the rapid calculation of derivatives. Backpropagation networks may be broken down into two primary categories:

1. **Static Backpropagation.** A network called static backpropagation was built so that it could map static inputs onto static outputs. Static backpropagation networks have the ability to address challenges associated with static categorization, such as optical character recognition (OCR).
2. **Recurring backpropagation.** Fixed-point learning is accomplished with the help of the recurrent backpropagation network. The activation from the recurrent backpropagation feeds forward until it reaches a predetermined value.

What is a backpropagation algorithm in a neural network?

- Backpropagation is a type of learning algorithm that is utilised by artificial neural networks. Its purpose is to compute a gradient descent with regard to the weight values of the various inputs. The systems are tuned by modifying connection weights in order to reduce the gap as much as possible between the desired and accomplished outputs of the system. This is done by comparing the desired outputs to the actual outputs of the system.
- The weights in the method are updated in a backwards fashion, from the output to the input, hence the name of the algorithm.

The following is a list of advantages that may be gained by utilising a backpropagation algorithm:

- It does not have any settings that can be tuned, the only thing that can be adjusted is the amount of inputs.
- It has a great degree of adaptability, it is very efficient and it does not call for any prior information of the network.
- This is a tried-and-true method that has a solid track record of success.
- It is simple to use, quick and straightforward to programme.
- Users are not required to master any specialised functionalities in order to use it.

The following is a list of the drawbacks associated with utilising a backpropagation algorithm:

- A matrix-based technique is preferred over a mini-batch approach by this system.
- Noise and inconsistencies can easily throw off the results of data mining.
- The performance is quite sensitive to the data that is input.
- Training requires a significant investment of both time and resources.

Features of Backpropagation:

1. This technique is known as the gradient descent method and it is implemented when a basic perceptron network is utilised with a differentiable unit.
2. in contrast to other networks, it has a unique method for determining the weights of nodes throughout the “learning” stage of the network’s development, which distinguishes it from other networks.
3. Training consists of the following three stages:
 - The feed-forward of input training pattern
 - The computation of the mistake and its subsequent backward propagation
 - A revised estimate of the weight.

Working of Backpropagation

Backpropagation’s Methodology Neural networks employ supervised learning to produce output vectors from input vectors that the network operates on and backpropagation is the method by which this is accomplished. It Does a comparison of the output that was created with the output that was requested and creates an error report if the result does not match the output vector that was generated. Finally, in order to provide the result you want, it modifies the weights in accordance with the bug report.

Backpropagation Algorithm:

Step 1: X-based inputs are received via the route that was preconnected.

Step 2: In the second step, the input is represented by making use of the real weights W. In most cases, the weights are selected at random.

Step 3: Compute the output of each neuron moving from the input layer to the hidden layer and then to the output layer. This is the third step.

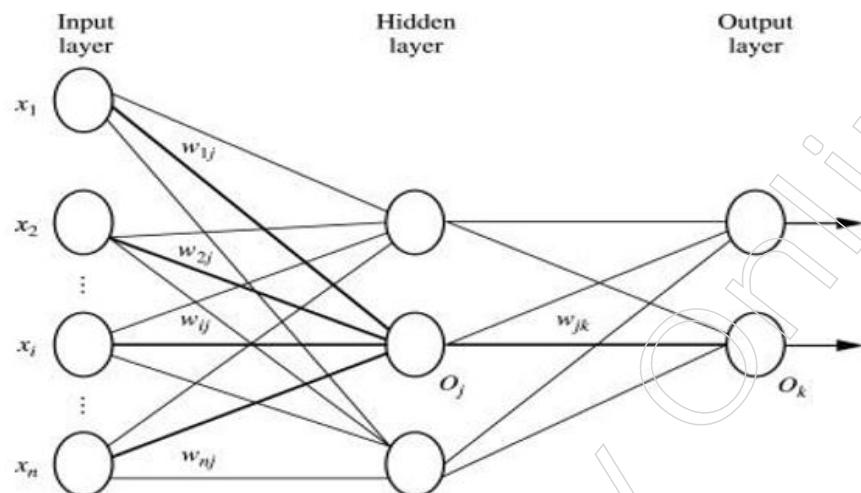
Step 4: Compute the amount of inaccuracy in the outputs at Step 4.

Backpropagation Error= Actual Output – Desired Output

Step 5: It involves going back to the hidden layer from the output layer in order to make adjustments to the weights in order to reduce the error.

Step 6: Do the procedure as many times as necessary until the desired results are obtained.

Notes



5.3.5 Neural Network in Data Science

The study of biological neural networks served as the inspiration for the development of artificial neural networks. These systems learn how to accomplish tasks by being shown with a variety of datasets and examples, but they are not given any rules that are particular to the jobs themselves. The general concept is that the system, which has not been pre-programmed with an awareness of the datasets it will be working with, will produce identifying features based on the data it has been given. The computational models for threshold logic serve as the foundation for neural networks. The concepts of algorithms and mathematics are brought together in threshold logic. Either the study of the human brain or the application of neural networks to the field of artificial intelligence serves as the foundation for neural networks. The work has contributed to advancements in the theory of finite automata. Neurons, the connections between them which are known as synapses, weights, biases, a propagation function and a learning algorithm are the components that make up a standard neural network. An input will be sent to neurons via previous neurons, each of which will have an activation, a threshold, an activation function f and an output function. Connections are made up of connections, weights and biases and they determine how one neuron sends its output to another neuron. In the process of propagation, both the input and the output are computed and the function of the preceding neurons is added to the weight. The term "learning" is used to indicate to the process of making changes to the neural network's free parameters, namely its weights and bias. The weights and thresholds of the variables in the network are altered as a direct result of the learning rule. The learning process may be broken down into three main stages, or sequences of occurrences. To name a few of these:

1. A model of the new environment is used to simulate the neural network.
2. Following that, the free parameters of the neural network are altered as a direct consequence of the simulation that was just run.
3. As a result of the adjustments made to its free parameters, the neural network reacts in an unexpected manner to the surrounding environment.

Supervised vs Unsupervised Learning: Neural networks learn through the process of supervised learning; In supervised machine learning, an input variable is denoted by

the letter x, while a desired output variable is denoted by the letter y. This is where we present the idea of a teacher who is knowledgeable about the surrounding environment. So, it is safe to assume that the instructor has both the input and the output set. The neural network does not take the surrounding environment into account. Both the instructor and the neural network have access to the input and the neural network produces an output depending on the input. After that, this output is compared with the output that the instructor has specified as being wanted and concurrently, an error signal is generated. The free parameters of the network are then modified one at a time in order to get the lowest possible error. When the algorithm reaches a level of performance that is deemed satisfactory, the learning process is terminated. The input data for unsupervised machine learning is denoted by X and there are no associated output variables. The purpose of this endeavour is to construct a model of the underlying structure of the data in order to gain a deeper knowledge of the data. Classification and regression are the terms that are most closely associated with supervised machine learning. Clustering and association are two of the most important concepts in unsupervised machine learning.

Evolution of Neural Networks: Hebbian learning examines neuronal plasticity and how it relates to the evolution of neural networks. Learning under the Hebbian model takes place in an unsupervised setting and focuses on long-term potentiation. Learning according to the Hebbian paradigm focuses on pattern recognition and exclusive-or circuits; it examines if-then principles. Backpropagation was able to overcome the problem of exclusive-or, which Hebbian learning was unable to do. This not only made multi-layer networks viable and efficient, but it also made them possible. In the event that an error was discovered, that issue was fixed by adjusting the weights at each node across all of the layers. Because of this, linear classifiers, support vector machines and max-pooling came into being. Both feedforward networks and recurrent neural networks are susceptible to the problem of disappearing gradients because of their use of back propagation. Deep learning refers to this type of learning. Simulations of biophysical processes and neurotrophic computing both make use of hardware-based design approaches. They are capable of large-scale component analysis and their convolutional processing generates a new type of neural computing with analogue. Back-propagation was also resolved for many-layered feedforward neural networks thanks to this solution. Convolutional networks are utilised for the purpose of alternating between convolutional layers and max-pooling layers with linked layers (either completely or sparsely connected) leading up to a final classification layer. The instruction is conducted without any preliminary unsupervised practise. Each filter is analogous to a weights vector that requires training in order to function properly. When working with both small and big neural networks, the shift variance has to have a guarantee attached to it. Development Networks are working to find a solution to this problem. Learning by error correction, learning based on memory and learning through competition are some of the other methods of education.

Types of Neural Networks

There are several varieties of deep neural networks and each one, depending on the application, carries with it a unique set of benefits and drawbacks. Examples include:

- Input, convolution, pooling, fully connected and output layers make up the five different types of layers that are found in convolutional neural networks

Notes

(CNNs). Each layer has a distinct function, such as activating, linking, or summarising the previous layer's information. Image categorization and object recognition have become more popular thanks to convolutional neural networks. Nevertheless, CNNs have also been utilised in other fields, including as the analysis of natural languages and weather forecasting.

- Sequential information, such as time-stamped data from a sensor device or a spoken sentence formed of a sequence of words, is input into recurrent neural networks (RNNs). A recurrent neural network differs from standard neural networks in that its inputs are not independent of one another. Instead, the output of each element of the network is reliant on the calculations performed by the components that came before it. RNNs are utilised in applications involving forecasting and time series, as well as applications involving sentiment analysis and other text-based tasks.
- Feedforward neural networks, in which every single perceptron in one layer of the network is linked to every single perceptron in the layer above it. The only way for information to go from one layer to the next is in a forward manner and that information is fed forward. There are not any feedback loops in this system.
- Autoencoder neural networks are utilised in the process of developing abstractions referred to as encoders, which are produced from a specified collection of inputs. Although autoencoders are quite similar to more typical neural networks, their goal is to model the inputs themselves; as a result, this technique is referred to as an unsupervised learning approach. The goal of autoencoders is to make relevant information more sensitive while decreasing sensitivity to non-relevant information. More abstractions are defined at higher tiers when new layers are added (layers closest to the point at which a decoder layer is introduced). Following that, either linear or nonlinear classifiers may utilise these abstractions as input.

5.4 Data Science and Ethical Issues

Introduction

Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data. This can be accomplished through the application of statistical modelling, scientific computing, scientific methods, processes and algorithms.

In addition to this, domain knowledge from the underlying application domain is included into data science (e.g., natural sciences, information technology and medicine). Data science is a multidimensional field that encompasses not only a science but also a research paradigm, a research technique, an academic discipline, a process and a professional occupation.

In order to “understand and analyse actual events” with the use of data, a “concept to unite statistics, data analysis and informatics and their related approaches” has been developed called “data science.”

Under the framework of mathematics, statistics, computer science, information science and domain knowledge, it makes use of techniques and theories borrowed

from a wide variety of subjects. The study of data, on the other hand, is distinct from computer science and information science. The recipient of the Turing Award, Jim Gray, asserted that “everything about science is changing because of the impact of information technology” and the data deluge. Gray envisioned data science as a “fourth paradigm” of science, which would follow empirical, theoretical, computational and now data-driven approaches to scientific inquiry.

5.4.1 Role of FAT in Data Science, Ethical Challenges in Data Science

The file allocation table, often known as FAT, is a type of file system that was designed specifically for use with hard drives. When it was first established, FAT employed either 12 or 16 bits for each cluster item in the file allocation table. The operating system (OS) relies on it to handle data on hard drives and other types of computer systems. Moreover, it is frequently discovered on flash memory, digital cameras and portable electronic gadgets. It is utilised to store information regarding files and to lengthen the life of a hard disc.

Seeking is a procedure that is required by most hard drives. Seeking refers to the actual process of physically looking for data and positioning the read/write head of the hard disc. The FAT file system was developed to cut down on the amount of searching that is required, which in turn reduces the amount of wear and tear that is placed on the hard disc.

The File Allocation Table (FAT) was developed to handle both hard discs and subdirectories. The older version of FAT12, known as FAT12, limited cluster addresses to 12-bit values and supported a maximum of 4078 clusters. When used with UNIX, this limitation was increased to a maximum of 4084 clusters. The more effective FAT16 extended to a 16-bit cluster address, which allowed for up to 65,517 clusters per volume. It also included 512-byte clusters with 32MB of capacity and a bigger file system; when combined with the four sectors, it was 2,048 bytes.

IBM was the first company to produce FAT16 in 1983, coinciding with the debut of IBM's personal computer AT (PC AT) and Microsoft's MS-DOS (disc operating system) 3.0 software. 1987 saw the introduction of Compaq DOS 3.31, which included an increase in the disc sector count to 32 bits and an expansion of the original FAT16 file system. Because the disc was intended to be used with a 16-bit assembly language, the disc as a whole has to be modified in order to employ sector numbers that are 32 bits in length.

FAT32 was initially released by Microsoft in 1997. The FAT file system's capacity restrictions were extended and it enabled DOS real mode code to manage the format. The cluster address in FAT32 is 32 bits and 28 of those bits are utilised to store the cluster number for a maximum of about 268 million FAT32 clusters. A partition is the greatest level of split that may occur within a file system. The partition is broken up into volumes, which are essentially logical hard discs. It is common practise to designate a letter, such as C, D, or E, for each logical drive.

A FAT file system is comprised of four distinct parts, each of which is represented as a separate structure within the FAT partition. The following are the four divisions:

Notes

- **Boot Sector:** This section of the disc is also referred to as the reserved sector and it may be found at the beginning of the disc. It contains the OS's required boot loader code in order to start a PC system, a partition table referred to as the master boot record (MBR) that describes how the drive is organised and the BIOS parameter block (BPB) that describes the physical outline of the data storage volume. All of these things are contained within it.
- **FAT Region:** This area typically has two versions of the File Allocation Table, one of which is used for redundancy testing and the other of which describes how the clusters are distributed over the file system.
- **Data Region:** This is the region that stores the data for the directory as well as any existing files. The vast majority of the division is taken up by it.
- **Root Directory Region:** This region is a directory table that provides information about the files and directories that are located in the root directory. It is only used with the FAT16 and FAT12 file systems and not with any of the other FAT file systems. It has a predetermined maximum capacity, which may be set up during the creation process. The root directory is normally stored in the data area of the FAT32 file system so that it may be enlarged if necessary.

Advantages

- **Storage:** Partition sizes of up to 8 terabytes can be supported
- **Scope:** It is the most outdated of the three different file systems that are compatible with the Windows operating system. Because of this, a significant number of companies who produce computer accessories, such as USB drives, Game Consoles and so on, employ it extensively in their products.
- **Compatibility with various Filesystems:** The FAT32 filesystem makes it easy to convert between other formats. Consequently, it is possible to convert a volume that uses the FAT32 file system to an NTFS disc without any of the data being lost in the process.
- **Compatibility:** The filesystem may be used with the vast majority of operating systems that are now available, including Linux, Windows and Macintosh.

Disadvantages

- Allows for the storage of files of up to 4 gigabytes in size.
- Provides an insignificant level of protection for the data that is saved. As a result, the data are susceptible to unethical manipulation.
- Is incapable of tolerating any errors.
- There is no support for encrypting individual files or folders individually.
- There is no support for recovering data if it is lost, in contrast to more recent file systems, which incorporate specific methods to make it easier to recover data in the event that it is lost (ex. NTFS has Journalling which could be used for speeding up the process of data recovery).
- Several companies who produce technology have discontinued support for it because newer filesystems that are considered superior have taken precedent.

Challenges in Data Science

1. Reinforcing human biases

According to a prediction made by Gartner ('Gartner Says Almost Half of CIOs Are Preparing to Use Artificial Intelligence,' 2020), by 2022, 85% of data science initiatives would give incorrect results owing to bias in either the data, the algorithms, or the teams responsible for managing them.

Data from the past is crunched by algorithms used in data science to forecast the future. The judgements that humans have taken in the past constitute the basis for the generation of data. If the algorithm is trained based only on previously collected data, this might result in some of these biases being included into the algorithms. The data and hypotheses that analysts choose to focus on can also influence the algorithms they develop since they may prioritise different aspects of the problem.

2. Lack of transparency

In the field of data science, algorithmic processes frequently take the form of a "black box," in which a model makes a prediction but does not elaborate on the reasoning behind that prediction. This definition fits a large number of recently developed machine learning techniques. By using black box solutions, it might be difficult for a company to comprehend and articulate the rationale behind a certain business decision. According to Andrews, who makes the observation, "Whether an AI system gives the proper response is not the sole problem... The executives need to be aware of the factors that contribute to its success and be able to explain the logic behind its failures.

3. Privacy

Confidentiality Protecting an individual's personal information has emerged as a primary concern in recent years. Many organisations keep private information on file, which leaves it vulnerable to hacking and other forms of abuse. Cambridge Analytica, a data analytics business that worked on Donald Trump's election campaign, used customers' Facebook data to influence their voting behaviour in the 2016 United States presidential election. The corporation was hired by Trump to assist with his election campaign.

A huge data breach resulted in the collection of information from 50 million Facebook user profiles by Cambridge Analytica. Because of this occurrence, ethical issues regarding the improper use of data were brought to light. There has been a rise in the number of data breaches that occur in every region of the world. Companies are now subject to rules and regulations, such as the General Data Protection Regulation (GDPR), which monitor the manner in which they retain sensitive data and how they utilise it.

4. Consent and Power

There is a lack of transparency among organisations regarding the data that they gather and how they utilise it to make decisions. Most online browsers and websites secretly store massive quantities of information about their users, often without the users' knowledge or consent. For instance, Google (Chrome and Gmail) and Facebook

Notes

Notes

keep the browsing data of individual users and make money off it by selling advertising insights derived from the consumers' data.

5.4.2 Some Real life Examples: Covid19, Data breach Cases

More than 1.2 billion individuals have been infected with COVID-19 (Coronavirus) and out of them, around 65,000 people have died as a direct result of this disease up till now. Fresh instances of COVID-19 (Coronavirus) are fast growing at startling rates globally. Because of the sudden increase in reported instances and the health information associated with them, an important new source of information and knowledge has been established. The immediate need to store such a vast volume of data in these situations using a variety of data storage technologies is a necessity that must be met immediately. These data are put to use in research and development projects pertaining to the virus, the pandemic and the actions that may be taken to combat this infection and its aftereffects. Big data refers to a new technology that enables the digital storage of a significant amount of information on these patients. A computational analysis can be helpful in revealing patterns, trends, correlations and differences. It may also aid in offering insights about the spread of this virus as well as the methods used to manage it. Big data, which has the power of gathering data in great detail, may be put to productive use to reduce the likelihood of this virus spreading.

By the use of big data technologies, it is possible to keep a vast quantity of information on the individuals who have been infected with the Covid-19 virus. It is useful in gaining a more in-depth grasp of the nature of this virus. The data that was collected may then be trained over and over again in order to develop new preventative strategies in the future. This technology is utilised to keep the data of all of the many sorts of cases affected by Covid-19, including those that are infected, recovered and expired. This information may be put to good use in identifying cases and helping to distribute resources for the purpose of improving public health protection. Numerous different kinds of digital data, including as a patient's location, proximity, patient-reported travel, co-morbidity, patient physiology and current symptoms, can be digitised and used for the purpose of creating actionable insights at both the community and the demographic level. Applications of big data that are particularly relevant to the COVID-19 pandemic are shown in Table 1.

Table 1 : Significant applications of big data in COVID-19 pandemic

S. No.	Applications	Description
1	Identification of infected cases	<p>It is capable of storing the complete medical history of all patients, due to its capability of storing a massive amount of data</p> <p>By providing the captured data, this technology helps in identification of the infected cases and undertake further analysis of the level of risks</p>

2	Travel history	Used to store the travel history of the people to analyse the risk Helps to identify people who may be in contact with the infected patient of this virus
3	Fever symptoms	Big data can keep the record of fever and other symptoms of a patient and suggest if medical attention is required Helps to identify the suspicious cases and other misinformation with the appropriate data
4	Identification of the virus at an early stage	Quickly helps to identify the infected patient at an early stage Helps to analyse and identify persons who can be infected by this virus in future
5	Identification and analysis of fast-moving disease	Helps to effectively analyse the fast-moving disease as efficiently as possible

Notes

Big data makes available a vast quantity of information to researchers, healthcare professionals and epidemiologists, which enables these professionals to make more educated decisions in their fight against the COVID-19 virus. This data may be put to use in the pursuit of continually tracking the virus on a global scale and in the pursuit of innovation in the field of medicine. It is possible to use it to anticipate the influence that COVID-19 will have on a specific location as well as on the entire population. It contributes to the advancement of research and the creation of novel therapeutic methods. Big data may also aid individuals by pointing them in the direction of potential resources and possibilities, which in turn makes it easier to deal with stressful situations. In general, this technology offers data that can be used to carry out an analysis of the disease transmission and movement system, as well as the health monitoring and preventive system.

Big data may assist in the predictive analysis of orthopaedic and trauma surgery, which is performed using the data that is already accessible. Orthopaedics is a subspecialty of the field of surgery known as trauma. Mortality and morbidity from trauma are similar to those seen in the present epidemic and it's possible that ideas can be cross-fertilized between the two. This technology is important in preserving records of orthopaedic patients, which can impact clinical decision-making in a favourable and proper way. So, it has the potential to improve performance by providing a better knowledge of the treatment methods based on the proper data that has been recorded. When it comes to conducting clinical trials, big data is a godsend since it helps to speed up the treatment process by analysing a patient's medical history.

The analytics of big data will serve as a medium for the purposes of tracking, regulating, researching and preventing COVID-19 from becoming a pandemic. Manufacturing will be diversified and the creation of vaccines will be enhanced through more comprehensive techniques, with absolute knowledge. For the purpose of predicting the COVID-19 cure and identifying the symptoms associated with it,

Notes

prevalent modelled data helps in understanding and offers an edge over the other process. For example, corresponding homology models predicted by fold and function assignment system server for each target protein were downloaded from Protein Data Bank. The use of big data enables the provision of insights and analysis into the elements that lead to improved containment of the infected patients. China was successful in containing COVID-19 by collecting data about it and using AI to put the plan into action, which resulted in a reduced incidence of disease transmission. There are many different aspects of this epidemic that include big data, including as biological research, natural language processing, mining the scientific literature and social media. AI has the potential to play a large role in all of these areas.

To be successful in the surgical speciality of orthopaedics, one has to possess superior comprehension, respectable levels of physical strength, outstanding surgical skills and clinical acuity. New technologies, such as artificial intelligence, have been implemented in recent years as a supplement to these criteria. This has helped to produce advances in the field of orthopaedics and has also given a good influence in the treatment and surgery of patients. Big data, artificial intelligence and 3D printing are examples of emerging technologies that have the potential to facilitate significant change and innovation. These technologies open the door to possibilities for improved patient care and outcomes.

In some locations, the big data contains information that may be used to detect cases of this virus that may have been present. It contributes to the provision of an effective method of preventing the sickness and helps to extract additional useful information. Big data will, in the not too distant future, be of assistance to the general people, as well as to doctors, other healthcare professionals and researchers, in their efforts to track this virus and investigate the infection mechanism of COVID-19. The data that are presented are helpful in examining ways in which this illness might be delayed or finally averted. They are also helpful in optimising the allocation of resources and helping, as a result, to making decisions that are suitable and timely. In addition, with the support of this digital data storage technology, medical professionals and scientists are able to create a COVID-19 testing procedure that is both practical and effective.

Data Breach Cases

1. Yahoo – 3,000,000,000 records lost



In 2013, malicious actors broke into Yahoo's computer system and stole information from more than 3 billion user accounts. It is to everyone's good fortune that the information that was taken did not include sensitive data such as bank account numbers, unhashed passwords, or payment information.

2. River City Media – 1,370,000,000 records lost



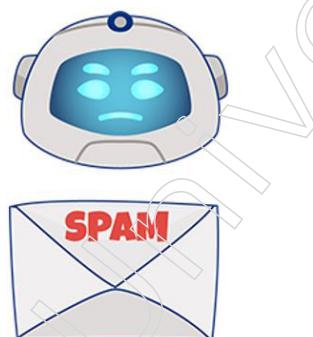
In March of 2017, a spam email operator made a mistake that resulted in the exposure of 1.37 billion records, making it one of the most significant data breaches in history. This breach occurred as a result of River City Media inadvertently publishing a snapshot of a backup from January 2017 without providing any kind of password security.

3. Aadhaar – 1,100,000,000 records lost



A state-owned utility firm in India suffered a compromise in March of 2018, which allowed unauthorised access to the biometric database known as Aadhaar. Because to this security compromise, each and every person of India who had their details registered was exposed, including their identity numbers, bank information and names. The compromised information was offered for sale on WhatsApp for a price of less than £6.

4. Spambot – 711,000,000 records lost



Due to a setup error, a spambot exposed users' passwords and email addresses in August of 2017. As a direct consequence of this, approximately 700 million data were compromised, which is roughly equivalent to one email address being disclosed for every man, woman and kid in Europe. Yet, this data breach featured a large number of duplicate and fictitious accounts.

5. Facebook – 533,000,000 records lost



Hackers attacked the social media behemoth Facebook in March 2019, taking advantage of a vulnerability that had been fixed the previous year, 2019. In one hacker forum, a staggering 533 million user records were uploaded, coming from 106 different nations. These details included the entire names and email addresses of users, in addition to their phone numbers, localities and biographical information.

Notes

6. Syniverse – 500,000,000 records lost



Syniverse, a company that is an essential component of the global telecommunications infrastructure, disclosed on September 27, 2021, in a filing with the United States Securities and Exchange Commission (SEC) the fact that hackers were able to access 500 million records within the company's database.

Syniverse is utilised by a large number of telecommunications firms all around the world, including AT&T, Verizon, T-Mobile, China Mobile and Vodafone. The information that was compromised included private data on the company's workers, as well as commercial secrets and intellectual property, as well as sensitive data about the company's clients, suppliers and other vendors, as well as other crucial financial data.

In addition, the business found out that hackers had been in their system for years, which means that the data breach may have possibly affected over 200 of its customers and millions of mobile users all over the world.

7. Yahoo – 500,000,000 records lost



In September of 2016, an actor working on behalf of a state stole 500 million records from Yahoo. These records included names, dates of birth and other sensitive information. This was the most significant breach of data security ever recorded at the time.

8. MySpace – 427,000,000 records lost



In May of 2016, a search engine that specialises in stolen data and a hacker stole more than 400 million records from MySpace. Both parties claimed that they had got the data via a previous data security issue that had not been disclosed to authorities. The information that was compromised included email addresses, passwords, usernames and even secondary passwords. On the dark web, the hacker offered to sell the information for \$2,800, which is equivalent to 6 Bitcoin.

9. Friend Finder Network – 412,000,000 records lost



In November of 2016, Friend Finder Network, a corporation that specialises in adult dating and entertainment, was attacked by cybercriminals. As a direct consequence of this, more than 412 million user accounts were made public. The hackers were also successful in leaking 339 million accounts from the website AdultFriendFinder.com. Among these accounts were 15 million "deleted" accounts that had never been removed from the server of the website.

10. Marriott International – 383,000,000 records lost



In September of 2018, fraudsters broke into the reservation system used by all Starwood hotels, including Westin, Le Meridien and Sheraton. As a result, Marriott International was forced to delete 383 million guest records from its database. They took personal information dating all the way back to 2014, including credit card numbers, passport information and other sensitive data.

Notes

Summary

- Text mining is the process of exploring and analysing huge volumes of unstructured text data with the assistance of software that can recognise concepts, patterns, subjects, keywords and other properties included within the data.
- Information Retrieval (IR) is a software programme that deals with the organisation, storage, retrieval and assessment of information from document repositories, particularly textual information.
- Hunting for potential threats includes conducting proactive investigations within a network to seek for irregularities that might point to a security breach.
- A data architect is the one who is accountable for the whole design, development, management and deployment of data architecture. This individual also determines how data is to be saved and accessed, while other decisions are made by internal bodies.
- Business requirements includes the aspects such as the growth of the company, the efficiency with which users can access the system, data management, transaction management and the utilisation of raw data through the transformation of said data into image files and records, which is followed by their storage in data warehouses. The primary part of storing commercial transactions is done through the usage of data warehouses.
- Business policies are the policies, which are essentially a set of rules, are helpful for explaining the manner in which data is processed. Internal organisational entities in addition to other government agencies are responsible for formulating these policies.

Glossary

- **Structured data:** This data is standardised into a tabular structure with several rows and columns, which makes it easier to store and process for analysis and machine learning algorithms.
- **Unstructured data:** This kind of data does not adhere to any particular data format that has already been established. Text from various sources, such as social media or product reviews, as well as rich media formats, such as video and audio files, may be included in this section.
- **Semi-structured data:** As the name implies, this type of data is a combination of structured and unstructured data formats. The data in this type can be organised in a variety of ways.

Notes

- **Julia:** It is an essential language for data science that aspires to be straightforward while retaining a high level of capability and has a syntax that is comparable to that of MATLAB or R.
- **Scala:** It has quickly risen to become one of the most popular languages for use cases involving artificial intelligence and data science.
- **Java:** It is a concurrent computer programming language that is class-based, object-oriented and was developed primarily to have as few implementation dependencies as possible

Check Your Understanding

1. What is the full form of NLP?
 - a) Natural Language Processing
 - b) Name Language Processing
 - c) Natural Language Preserving
 - d) No Language Processing
2. What is the full form of MIS?
 - a) Management Information System
 - b) Managed Information System
 - c) Model Information System
 - d) Machine Information System
3. The act of automatically creating a condensed version of a certain text that contains useful information for the end-user is referred to as _____.
 - a) Text summarisation
 - b) Clustering
 - c) Categorisation
 - d) Retrieval
4. What is the full form of SQL?
 - a) Stored Query Language
 - b) Simple Query Language
 - c) Structured Query Language
 - d) Semi – Structured Query Language
5. _____ is a computer language for statistics that is frequently utilised for statistical analysis, data visualisation and several other types of data manipulation.
 - a) C
 - b) C++
 - c) R
 - d) OOPS

6. What is the full form of SIEM?
 - a) Security Information and Event Management
 - b) Security Internet and Event Management
 - c) Severe Information and Event Management
 - d) Security Information and Event Modelling
7. What is the full form of EDR?
 - a) Endpoint Division and Response
 - b) Endpoint Detection and Recognition
 - c) Endpoint Detection and Response
 - d) Endpoint Detection and Risk
8. What is the full form of DNS?
 - a) Domain Name System
 - b) Division Name System
 - c) Direct Name System
 - d) Diligent Name System
9. What is the full form of ATT&CK framework?
 - a) Application Tactics Techniques and Common Knowledge
 - b) Adversary Tactics Techniques and Common know-how
 - c) Adversary Tactics Techniques and Common Knowledge
 - d) Advisory Tactics Techniques and Common Know-how
10. A _____ called a wireless logger, also known as a wireless sensor, is a form of data logger that retrieves data through the use of wireless technology (such a mobile app or Bluetooth) and then sends that data using cloud technology.
 - a) Standalone Data Logger
 - b) Wireless data logger
 - c) Computer-based data logger
 - d) Web-based data logger
11. The act of importing data from websites into files or spreadsheets is referred to as "web scraping" and is also known as _____.
 - a) Data scraping
 - b) HTML Parsing
 - c) DOM Parsing
 - d) Vertical Aggregation
12. What is the full form of DOM?
 - a) Document Object Model

Notes

- b) Dominant Object Model
 - c) Document Oriented Model
 - d) Document Obtained Model
13. _____ are organised in a tree-like form, scrapers may utilise XPath to browse through XML documents by picking nodes based on a variety of criteria.
- a) XML documents
 - b) Google Sheets
 - c) DOM Parsing
 - d) HTML Parsing
14. _____ is the process of fixing or removing wrong, corrupted, incorrectly formatted, duplicate, or incomplete data from a dataset.
- a) Data-artifacts
 - b) Data Compatibility
 - c) Data cleaning
 - d) Data Validation
15. The process of arranging data in such a way that it is simpler to understand, more designed, or more organised is referred to as _____.
- a) Data Validation
 - b) Data Cleaning
 - c) Data Optimisation
 - d) Data Manipulation
16. The values or data that are not saved (or are not available) for some variable/s in the provided dataset are referred to as _____.
- a) Missing Data
 - b) Validate Data
 - c) Clean Data
 - d) Duplicate Data
17. What is the full form of MCAR?
- a) Managed Completely at Random
 - b) Missing Completely at Random
 - c) Missing Common at Random
 - d) Managed Common at Random
18. What is the full form of ETL?
- a) Evaluate, Transform, Load
 - b) Extract, Transform, Load

- c) Extract, Transact, Load
d) Extract, Transform, Leverage
19. _____ is a term that refers to the methodology, tools and applications that are used to collect, analyse and derive insights from a wide variety of data sets that move at a high volume and velocity.
- a) Software Artifacts
b) Documentation Artifacts
c) Big data analytics
d) Data Science
20. _____ is a type of business model that makes use of the Entity Relationship (ER) model to determine the relation that exists between entities and the attributes of those entities.
- a) Conceptual Model
b) Logical Model
c) Physical Model
d) Data Model

Notes**Exercise**

1. Write a short note on text mining.
2. Explain the concept of Data Science.
3. List various data science programming languages.
4. Explain the concept of Data cleaning.
5. What are Outlines. Explain its types.
6. Explain four V's of Big Data.

Learning Activities

1. Explain the concept of Data-Hunting, Logging, Scrapping.
2. Explain the concept of Big Data Analytics.
3. Explain Data architecture design and explain the difference between Human Brain and Artificial Network.

Check Your Understanding – Answers

1. a)
2. a)
3. a)
4. c)
5. c)
6. a)

Notes

7. c)
8. a)
9. c)
10. a)
11. a)
12. a)
13. a)
14. c)
15. d)
16. a)
17. b)
18. b)
19. c)
20. a)

Further Readings and Bibliography

1. <https://www.techopedia.com/definition/1369/file-allocation-table-fat>
2. <https://www.geeksforgeeks.org/fat-full-form/>
3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204193/>
4. <https://termly.io/resources/articles/biggest-data-breaches/>

