



Data Analysis and Visualisation

Study Guide



Table of Contents

Welcome.....	1
Introduction to Business Analytics.....	3
1.1. Welcome to the Introduction to Business Analytics Unit.....	3
1.1.1. Introduction to this Unit	3
1.2. KT0101 - Overview	4
1.3. KT0102 - Business decisions and analytics.....	5
1.4. KT0103 - Types of business analytics.....	6
1.5. KT0104 - Applications of business analytics	7
1.6. KT0105 - Data science overview	9
Introduction to business processes, analysis and process modelling	10
2.1. Welcome to the Introduction to business processes, analysis and process modelling Unit.....	11
2.1.1. Introduction to this Unit	11
2.2. KT0201 - Business analysis.....	12
2.3. KT0202 - Business processes.....	13
2.4. KT0203 - Core business processes	14
2.5. KT0204 - Business process management.....	15
2.6. KT0205 – Types of business operations.....	17
2.7. KT0206 - Business process modelling notation (BPMN).....	20
2.8. KT0207 - Difference between process and workflow.....	21
2.9. KT0208 - Workflow concepts, types and differences	24
2.10. KT0209 - Workflow-based logic	26
2.11. KT0210 - Workflow tools.....	27
2.12. KT0211 - Workflow based solutions or functionality.....	28
2.13. KT0212 - Typical major workflows to streamline and automate.....	29

Introduction to Data Science Programs	31
3.1. Welcome to the Introduction to Data Science Programs Unit	32
3.1.1. Introduction to this Unit	32
3.2. KT0301 - The three most popular languages in data science	33
3.3. KT0302 - Underlying design philosophy, grammar and data structures	34
3.4. KT0303 - Importance of programming and programming languages	35
3.5. KT0304 - Productivity (e.g. GitHub, git, Unix/Linux, and RStudio).....	37
Data Analytics	38
4.1. Welcome to the Data Analytics Unit.....	39
4.1.1. Introduction to this Unit	39
4.2. KT0401 - Data types and variables.....	40
4.3. KT0402 - Operators.....	41
4.4. KT0403 - Conditional statements.....	42
4.5. KT0404 - Loops.....	43
4.6. KT0405 - Script	44
4.7. KT0406 - Functions.....	45
4.8. KT0407 - Probability.....	46
4.9. KT0408 - Inference and modelling	47
Wrangling.....	48
5.1. Welcome to the Wrangling Unit	49
5.1.1. Introduction to this Unit	49
5.2. KT0501 - Importing data from different file formats.....	50
5.3. KT0502 - Web scraping	52
5.4. KT0503 - How to tidy data using suitable software packages to better facilitate analysis	54
5.5. KT0504 - String processing with regular expressions (regex).....	55
5.6. KT0505 - HTML parsing	56
5.7. KT0506 - Wrangling data using suitable software package	57
5.8. KT0507 - How to work with dates and times as file formats.....	59
5.9. KT0508 - Text mining	61

Data Structures	62
6.1. Welcome to the Data Structures Unit	63
6.1.1. Introduction to this Unit	63
6.2. KT0601 - Introduction to data structures	64
6.3. KT0602 - Identifying data structures.....	65
6.4. KT0603 - Assigning values to data structures	66
6.5. KT0604 - Data manipulation	67
Data Visualization	69
7.1. Welcome to the Data Visualization Unit.....	70
7.1.1. Introduction to this Unit	70
7.2. KT0701 - Introduction to data visualization.....	71
7.3. KT0702 - Data visualization using graphics	72
7.4. KT0703 - Data visualization using system for declaratively creating graphics	74
7.5. KT0704 - File formats of graphic outputs	75
High-throughput	76
8.1. Welcome to the High-throughput Unit.....	77
8.1.1. Introduction to this Unit	77
8.2. KT0801 - Organizing high throughput data.....	78
8.3. KT0802 - Multiple comparison problem	79
8.4. KT0803 - Family wide error rates.....	80
8.5. KT0804 - False discovery rate	82
8.6. KT0805 - Error rate control procedures.....	83
8.7. KT0806 - Bonferroni correction	84
8.8. KT0807 - q-values.....	85
8.9. KT0808 - Statistical modelling.....	86
8.10. KT0809 - Hierarchical Models and the basics of Bayesian Statistics	88
8.11. KT0810 - Exploratory data analysis for high throughput data.....	89
High-dimensional data analysis	90
9.1. Welcome to the High-dimensional data analysis Unit.....	91

9.1.1.	Introduction to this Unit	91
9.2.	KT0901 - Mathematical distance	92
9.3.	KT0902 - Dimension reduction	93
9.4.	KT0903 - Singular value decomposition and principal component analysis.....	94
9.5.	KT0904 - Multiple dimensional scaling plots	96
9.6.	KT0905 - Factor analysis.....	97
9.7.	KT0906 - Dealing with batch effects	98
9.8.	KT0907 - Clustering	99
9.9.	KT0908 - Heatmaps.....	100
	Basic machine learning and artificial intelligence concepts	101
10.1.	Welcome to the Basic machine learning and artificial intelligence concepts Unit.....	102
10.1.1.	Introduction to this Unit	102
10.2.	KT1001 - ML concepts and principles	103
10.3.	KT1002 - ML application	104
10.4.	KT1003 - ML technologies.....	107
10.5.	KT1004 - Supervised learning.....	109
10.6.	KT1005 - Unsupervised learning	110
10.7.	KT1006 - Reinforcement learning	111
10.8.	KT1007 - Algorithms.....	112
	Content Sources.....	113

Welcome

Welcome

Welcome to an exciting journey of discovery and learning at the Academic Institute of Excellence (AIE). AIE is a revolutionised family of brands, people, and students, offering a curriculum modelled on workplace scenarios to provide you with a practical understanding of how to apply your skills in the real world.

At AIE, we aim to provide innovative, quality education, coupled with excellent service delivery and a modern outlook on learning technology. We take pride in developing and supporting students through innovative programs that create well-rounded, employable, and professionally developed individuals. Our learning paths are meticulously designed based on global skill demands and the needs of modern students. Our facilitators are qualified subject matter experts with practical industry experience and tertiary education expertise.

Conveniently located in Greenside, Midrand (Johannesburg), and De Waterkant (Cape Town), our campuses offer a range of recreational possibilities and ample workspace, including computer studios, libraries, and a fully operational Makers Lab. You'll have access to lecturers, resources, and the latest software to help you fulfill your course requirements.

At AIE, we are committed to delivering demand-driven education built upon the principles of quality education, innovation, and technology. Each of our programs is benchmarked against local and international standards to ensure you receive the highest level of quality education.

Join us as we strive towards excellence and empower future generations to become problem solvers, critical thinkers, and innovators—the leaders of tomorrow.

Our VISION

To deliver demand-driven education, built upon the principle of quality education through innovation and technology.



#Flexibility



#AnywhereAnytime



#Accessibility



#StudentSupport

Unit Name 1

Unit Name 2

Unit Name 3

Unit Name 4

Unit Name 5

Next

Data Analysis and Visualisation **DAAV****Module Purpose**

This module aims to provide a foundational understanding of business analytics, data science, and the various methodologies employed in analyzing and visualizing data. From introductory concepts in business processes to advanced techniques in handling high-dimensional data, this module covers a wide array of topics to prepare you for the challenges and opportunities in the field of data analytics.

Specific Outcome	Assessment Criteria
KM-08-KT01: Introduction to Business Analytics	Gain an overview of business analytics, understanding its importance in deriving insights from data to inform business decisions.
KM-08-KT02: Introduction to Business Processes, Analysis, and Process Modeling	Explore the fundamentals of business processes and how to model and analyze these processes for optimization and efficiency.
KM-08-KT03: Introduction to Data Science Programs	Learn about the various data science programs and tools available for conducting sophisticated data analysis.
KM-08-KT04: Data Analytics	Dive deep into data analytics techniques, including statistical analysis, predictive modeling, and exploratory data analysis.
KM-08-KT05: Wrangling	Master the art of data wrangling, preparing and cleaning data to make it suitable for analysis.

KM-08-KT06: Data Structures	Understand the different data structures used in data analysis and how they influence the analysis process.
KM-08-KT07: Data Visualization	Learn the principles of data visualization and how to use visual elements to communicate data insights effectively.
KM-08-KT08: High-throughput	Explore techniques for managing and analyzing high-throughput data sets, crucial in big data scenarios.
KM-08-KT09: High-dimensional Data Analysis	Gain insights into the challenges and strategies for analyzing high-dimensional data, including dimensionality reduction techniques.
KM-08-KT10: Basic Machine Learning and Artificial Intelligence Concepts	Get introduced to the basics of machine learning and AI, understanding their application in data analysis and prediction.



COMPUTER Requirements

It is advisable that students make use of their own personal computers to complete this module.



READ – SAQA Qualification Detail

This module forms part of the Occupational Certificate: Data Science Practitioner Occupational Certificate 118708, NQF 5 Follow the link on AMI to view the full details of the qualification.

Read: Data Science Practitioner SAQA Qualification

Link to SAQA document: [Click here](#)

Introduction to Business Analytics

Unit 1

Unit Overview

The following topics are covered in this unit:

- Overview
- Business decisions and analytics
- Types of business analytics
- Applications of business analytics
- Data science overview

Learning Outcomes

At the end of this unit the student should be able to

- Gain an overview of business analytics, understanding its importance in deriving insights from data to inform business decisions.

1.1. Welcome to the Introduction to Business Analytics Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

1.1.1. Introduction to this Unit

Business analytics is the practice of using data-driven techniques to inform and improve business decisions. It bridges the gap between raw data and actionable insights, enabling organizations to make informed choices and drive strategic outcomes. At its core, business analytics involves collecting, processing, and analyzing data to identify trends, patterns, and correlations. This process empowers businesses to make better predictions, optimize operations, and enhance customer experiences, ultimately leading to improved performance and competitive advantage.

Business analytics is pivotal in decision-making, as it helps organizations move from intuition-based decisions to evidence-based strategies. It encompasses three main types: descriptive analytics, which provides insights into past performance; predictive analytics, which forecasts future outcomes based on patterns and trends; and prescriptive analytics, which recommends optimal actions to achieve desired goals. These approaches are applied across various domains, such as marketing, finance, supply chain management, and human resources, to solve problems, identify opportunities, and streamline operations.

As part of the broader field of data science, business analytics utilizes statistical methods, machine learning algorithms, and visualization tools to transform data into actionable knowledge. It serves as a foundational element of data science, focusing on interpreting and applying data to address real-world business challenges. Through its diverse applications—from optimizing inventory management to enhancing customer retention—business analytics plays a critical role in driving data-informed decisions in modern organizations.

Unit 1 – Introduction to Business Analytics

1.2. KT0101 - Overview

What is Business Analytics?

Business analytics, a data management solution and business intelligence subset, refers to the use of methodologies such as data mining, predictive analytics, and statistical analysis in order to analyze and transform data into useful information, identify and anticipate trends and outcomes, and ultimately make smarter, data-driven business decisions.

The main components of a typical business analytics dashboard include:

- **Data Aggregation:** prior to analysis, data must first be gathered, organized, and filtered, either through volunteered data or transactional records
- **Data Mining:** data mining for business analytics sorts through large datasets using databases, statistics, and machine learning to identify trends and establish relationships
- **Association and Sequence Identification:** the identification of predictable actions that are performed in association with other actions or sequentially
- **Text Mining:** explores and organizes large, unstructured text datasets for the purpose of qualitative and quantitative analysis
- **Forecasting:** analyzes historical data from a specific period in order to make informed estimates that are predictive in determining future events or behaviors
- **Predictive Analytics:** predictive business analytics uses a variety of statistical techniques to create predictive models, which extract information from datasets, identify patterns, and provide a predictive score for an array of organizational outcomes
- **Optimization:** once trends have been identified and predictions have been made, businesses can engage simulation techniques to test out best-case scenarios

- **Data Visualization:** provides visual representations such as charts and graphs for easy and quick data analysis

1.3. KT0102 - Business decisions and analytics

What is business analytics and decision-making?

Business analytics aims to generate knowledge, understanding and learning – collectively referred to as 'insight' – to support evidence-based decision making and performance management.

Why is business analytics important for decision-making?

Business analytics can help companies make better, more informed decisions and achieve a variety of goals. By leveraging data, businesses can: Better understand consumer behavior. Gain insight into their competitors.

1.4. KT0103 - Types of business analytics

Types of business analytics

There are four main types of business analytics companies can employ to better understand and grow their business.

- **Descriptive analytics**—is one of the most basic forms of analytics, providing insights on what has happened or is currently happening. Sales reports and social media engagement are examples of descriptive analytics.
- **Predictive analytics**—relies on tools like machine learning and artificial intelligence (AI) algorithms to project what will happen, such as how a product will sell or who will buy it.
- **Prescriptive analytics**—uses a variety of data points, such as available resources and past performance, to suggest a course of action or strategy to achieve a desired result.
- **Diagnostic analytics**—examines data to explain why something happened.

1.5. KT0104 - Applications of business analytics

What is the application of business analytics?

Business Analytics can help you in supply chain management, inventory management, measure performance of targets, risk mitigation plans, improve efficiency in the basis of product data, etc. For example: The Manager wants information on performance of a machinery which has been used past 10 years

6 Applications of Business Analytics with Examples:

- Finance

BA is of utmost importance to the finance sector. Data Scientists are in high demand in investment banking, portfolio management, financial planning, budgeting, forecasting, etc.

For example: Companies these days have a large amount of financial data. Use of intelligent Business Analytics tools can help use this data to determine the products' prices. Also, on the basis of historical information Business Analysts can study the trends on the performance of a particular stock and advise the client on whether to retain it or sell it.

- Marketing

Studying buying patterns of consumer behaviour, analysing trends, help in identifying the target audience, employing advertising techniques that can appeal to the consumers, forecast supply requirements, etc.

For example: Use Business Analytics to gauge the effectiveness and impact of a marketing strategy on the customers. Data can be used to build loyal customers by giving them exactly what they want as per their specifications.

- HR Professionals

HR professionals can make use of data to find information about educational background of high performing candidates, employee attrition rate, number of years of service of employees, age, gender, etc. This information can play a pivotal role in the selection procedure of a candidate.

For example: HR manager can predict the employee retention rate on the basis of data given by Business Analytics.

- CRM

Business Analytics helps one analyse the key performance indicators, which further helps in decision making and make strategies to boost the relationship with the consumers. The demographics, and data about other socio-economic factors, purchasing patterns, lifestyle, etc., are of prime importance to the CRM department.

For example: The company wants to improve its service in a particular geographical segment. With data analytics, one can predict the customer's preferences in that particular segment, what appeals to them, and accordingly improve relations with customers.

- Manufacturing

Business Analytics can help you in supply chain management, inventory management, measure performance of targets, risk mitigation plans, improve efficiency in the basis of product data, etc.

For example: The Manager wants information on performance of a machinery which has been used past 10 years. The historical data will help evaluate the performance of the machinery and decide whether costs of maintaining the machine will exceed the cost of buying a new machinery.

- **Credit Card Companies**

Credit card transactions of a customer can determine many factors: financial health, life style, preferences of purchases, behavioral trends, etc.

For example: Credit card companies can help the retail sector by locating the target audience. According to the transactions reports, retail companies can predict the choices of the consumers, their spending pattern, preference over buying competitor's products, etc. This historical as well as real-time information helps them direct their marketing strategies in such a way that it hits the dart and reaches the right audience.

Why analytics is important

Business analytics can help companies make better, more informed decisions and achieve a variety of goals. By leveraging data, businesses can:

- Better understand consumer behavior
- Gain insight into their competitors
- Identify market trends
- Measure accomplishments against goals
- Optimize operations

Business analytics has helped many companies navigate tough times, especially regarding the COVID-19 pandemic. According to a survey by business intelligence company Sisense, 50% of companies reported using analytics "more often" or "much more often," during the pandemic. The increased use of analytics was even more pronounced among smaller companies. Helping organizations navigate through crises is just one of the

many reasons why analytics is important to business. Data suggest companies that use analytics to their advantage are twice as likely to rank in the top quarter for financial performance, five times more likely to make timely decisions and three times more likely to execute their decisions and plans.

1.6. KT0105 - Data science overview

What is data science purpose?

The answer, in a nutshell, is simple: The purpose of data science is to find patterns. Understanding patterns means understanding the world. In everything, from a mechanic fixing a car to a scientist making a research breakthrough, identifying a pattern is the first step towards progress.

The Data Science Lifecycle

Data science's lifecycle consists of five distinct stages, each with its own tasks:

- **Capture:** Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.
- **Maintain:** Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.
- **Process:** Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.
- **Analyze:** Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.
- **Communicate:** Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the

Introduction to business processes, analysis and process modelling

Unit 2

Unit Overview

The following topics are covered in this unit:

- Business analysis
- Business processes
- Core business processes
- Business process management
- Types of business operations
- Business process modelling notation (BPMN)
- Difference between process and workflow
- Workflow concepts, types and differences
- Workflow-based logic
- Workflow tools

Learning Outcomes

At the end of this unit the student should be able to

- Explore the fundamentals of business processes and how to model and analyze these processes for optimization and efficiency.

Unit 2 – Introduction to business processes, analysis and process modelling

2.1. Welcome to the Introduction to business processes, analysis and process modelling Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

2.1.1. Introduction to this Unit

Business processes form the backbone of organizational operations, defining how tasks are structured, executed, and optimized to achieve business objectives. Business analysis plays a pivotal role in understanding these processes, identifying inefficiencies, and recommending improvements. Core business processes, such as procurement, production, and customer service, are essential for value creation and operational excellence. Business Process Management (BPM) provides a structured

approach to analyze, improve, and automate these processes to enhance efficiency and adaptability in a dynamic business environment.

A key tool in business process analysis is Business Process Modelling Notation (BPMN), a standardized graphical language that visually represents processes, enabling better communication among stakeholders. BPMN distinguishes between processes, which are sequences of activities aimed at achieving a goal, and workflows, which focus on task execution and coordination within a process. Understanding workflow concepts, such as sequential and parallel workflows, is crucial for designing efficient systems. Workflow-based logic underpins the execution of automated tasks, ensuring seamless transitions and decision-making within processes.

Organizations leverage workflow tools to streamline operations, improve collaboration, and reduce errors. These tools enable the design, execution, and monitoring of workflows, providing flexibility to accommodate varying business needs. The differences between workflows—manual, automated, and hybrid—highlight the importance of tailoring solutions to specific operational contexts. By integrating workflows into broader business process management frameworks, businesses can achieve a holistic approach to efficiency and innovation, ensuring continuous improvement and alignment with strategic goals.

Unit 2 – Introduction to business processes, analysis and process modelling

2.2. KT0201 - Business analysis

What is Business Analysis?

Business analysis (BA) is the practice of helping someone to solve a problem by delivering a solution that matches their needs. Common tasks associated with BA include advising strategic business options, improving business systems around specific outcomes, and defining IT system requirements.

Business analysis is a professional discipline of identifying business needs and determining solutions to business problems. Solutions often include a software-systems development component, but may also consist of process improvements, organizational change or strategic planning and policy development.

Business analysis is a professional discipline of identifying business needs and determining solutions to business problems. Solutions often include a software-systems development component, but may also consist of process improvements, organizational change or strategic planning and policy development. The person who carries out this task is called a business analyst or BA.

- Business analysts do not work solely on developing software systems. But work across the organisation, solving business problems in consultation with business stakeholders. Whilst most of the work that business analysts do today relate to software development/solutions, this derives from the ongoing massive changes businesses all over the world are experiencing in their attempts to digitise
- Although there are different role definitions, depending upon the organization, there does seem to be an area of common ground

where most business analysts work. The responsibilities appear to be

- To investigate business systems, taking a holistic view of the situation. This may include examining elements of the organisation structures and staff development issues as well as current processes and IT systems.
- To evaluate actions to improve the operation of a business system. Again, this may require an examination of organisational structure and staff development needs, to ensure that they are in line with any proposed process redesign and IT system development.

Unit 2 – Introduction to business processes, analysis and process modelling

2.3. KT0202 - Business processes

What is a Business Process?

A business process is defined as a collection of business tasks and activities that when performed by people or systems in a structured course, produce an outcome that contributes to the business goals. A business process includes at least one of, but not limited to, the following elements:

- task/ activity
- system
- employee(s)
- workflow
- data

Business processes are invented to derive and contribute to organizational goals. The continuous and repeated execution of business processes is pivotal to successful business operations and business growth.

Business process structures can be simple or complex, based on the elements involved in the process. Through every business process, a business strives to achieve certain goals.

Some key attributes that distinguish business processes from other business tasks and activities are:

- a process is repeatable
- a process is flexible and not rigid
- a process is specific and has established start and endpoints
- a process is measurable

To understand this better, let us take an example of the employee off boarding process, which has the following steps:

- acknowledging the resignation/termination
- negotiation with the employee for retainment
- notice period and final settlement procedures
- planning for pending projects and assigned tasks
- hand over and knowledge transfer of project and company details
- exit interview

These are the most common activities involved globally in an employee offboarding process. Through acknowledgment and negotiation discussions, the business tries to reason with and retain the employee. The notice period and final settlement procedures help the employee and company get clarity on the terms of the contract pending after the resignation. The project planning and handover activities help the business make sure that daily processes are not affected due to the resignation of the employee. Finally, the exit interview provides the business an opportunity to improve with detailed feedback on the company operations and policies.

2.4. KT0203 - Core business processes

What Is a Core Business Process?

Core business processes are also known as operational, primary, or essential processes. These terms all refer to the actions, capabilities, and activities that a business needs to create and deliver a product or service.

Though core business processes do many things for a company, the primary objective is to deliver value to customers or business partners.

The core processes need to work together to achieve a common goal in the most efficient way to maximize profits.

It's necessary to understand that a company's ability to identify and manage its core business processes efficiently is critical to success. If one core process fails, it puts the entire organization at risk.

Core Business Process vs. Core Business Function

Before we press on, we wanted to point out some key distinctions between core business processes and core business functions.

While they sound interchangeable, processes and functions are two different approaches to organization, coordination, optimization, and planning.

Differentiating between business processes and business functions across these four domains is helpful because it explains the foundation of most companies.

After all, how can you identify, refine, or optimize your business if you don't know which model to adopt?

Each approach has pros and cons, so determining which model to choose depends on your company's goals.

The primary difference between these methods is independence (function-driven) versus cohesiveness (process-driven), but like most things in business, it's not that simple.

2.5. KT0204 - Business process management

What is Business Process Management (BPM)?

Business process management (BPM), as defined by Gartner, is a discipline that uses various tools and methods to design, model, execute, monitor, and optimize business processes. A business process coordinates the behavior of people, systems, information, and things to produce business outcomes in support of a business strategy.

BPM focuses on putting a consistent, automated process in place for routine transactions and human interactions. It helps to reduce the business's operational costs by decreasing waste and rework, and by increasing the overall efficiency of the team.

Organizations engaged in BPM can choose to follow one of the various BPM methodologies, which include Six Sigma and Lean.

What BPM is Not

BPM is not a software product. There are BPM tools available that help in implementing standard and automated business processes. For example, HappyFox Workflows helps businesses automate complex, multi-step, and repetitive business processes. BPM, however, is not a software product in itself.

BPM is not Task Management. Task or Project Management is about handling or organizing a set of activities. A project management software like Microsoft Project, Jira, Asana, or Trello helps in managing tasks and ad-hoc projects. Business Process Management, on the other hand, is focused more on repetitive and ongoing processes that follow a predictable pattern or process management.

What are the Various Types of Business Process Management Systems?

BPM systems can be categorized based on the purpose that they serve. Here are the three types of business process management:

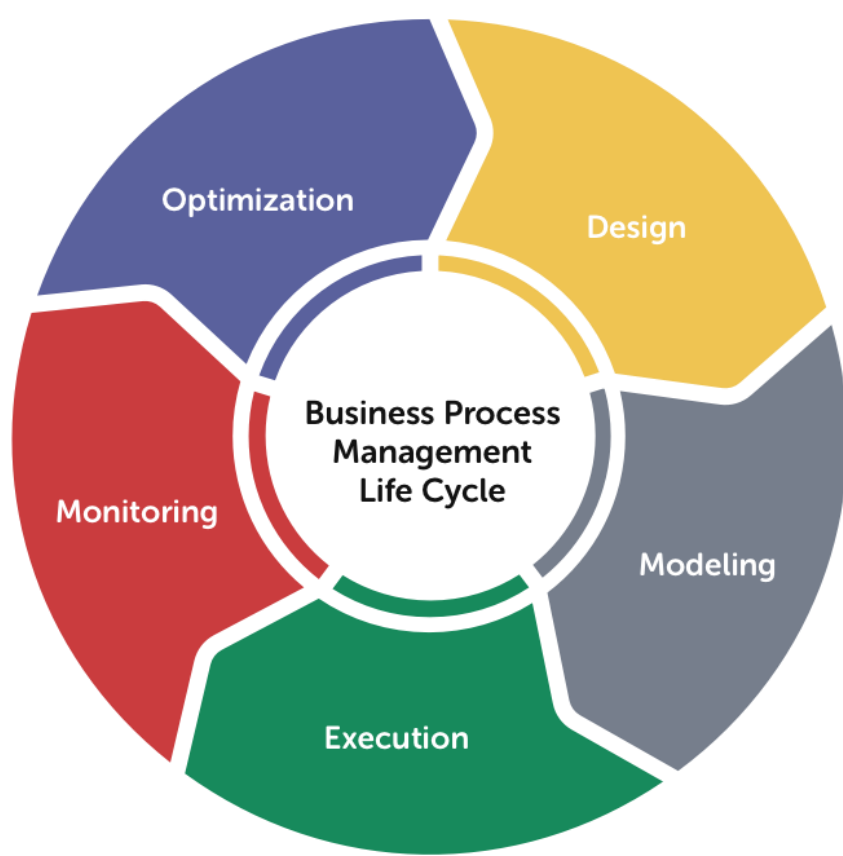
System-Centric BPM (or Integration-Centric BPM)

This type of business process management system handles processes that primarily depend on existing business systems (e.g., HRMS, CRM, ERP) without much human involvement. A system-centric business process management software has extensive integrations and API access to be able to create fast and efficient business processes. An example of an integration-centric process is online banking, which can include different software systems coming together.

Human-Centric BPM

Human-centric BPM considers the people first, supported by various automation functions. These are processes that are primarily executed by humans, and automation does not easily replace them. These often have a lot of approvals and tasks performed by individuals. Examples of human-centric processes include providing customer service, handling complaints, on-boarding employees, conducting e-commerce activities, and filing expense reports.

BPM Lifecycle: The 5 Steps in Business Process Management



Step 1: Design

Business analysts review current business rules, interview the various stakeholders, and discuss desired outcomes with management. The goal of the process design stage is to gain an understanding of the business rules and ensure if the results are in alignment with the organizational goals.

Step 2: Model

Modeling refers to identifying, defining, and making a representation of new processes to support the current business rules for various stakeholders.

Step 3: Execute

Execute the business process by testing it live with a small group of users first and then open it up to all users. In the case of automated workflows, artificially throttle the process to minimize errors.

Step 4: Monitor

Establish Key Performance Indicators (KPIs) and track metrics against them using reports or dashboards. It's essential to focus on the macro or micro indicators – an entire process vs. process segments.

Step 5: Optimize

With an effective reporting system in place, an organization can effectively steer operations toward optimization or process improvement. Business Process Optimization (BPO) is the redesign of the business processes to streamline and improve process efficiency and strengthen the alignment of individual business processes with a comprehensive strategy.

2.6. KT0205 – Types of business operations

Types of Business Processes

A collection of processes is what makes the business. Different processes come together to form an efficient business model. For a business to flourish it is important that the processes are optimized and the function smoothly in the organization. The primary function of processes is to add value to the business. There are different types of processes which are used to enhance the business. Basically, there are three categories under which all the types of business processes fall under.

A) Primary Types of Business processes :

These types of business process are extremely important and fundamental for business. They deal with the basic values and work alongside the vision and mission of the business. As the name suggests the primary process is a very basic process through which the business ensures delivery of services or products to their clients.

These processes are paid close attention to since they are an integral part of the organization. The ultimate aim of primary processes is to optimize themselves so that they add value at every step of the business as well as to the end product which internal value to the client. This improves the processes and has a direct effect on business output.

There are different types of business processes under primary processes. These are as follows:

1) Sales:

Types of Business Processes - 1

Sales standalone type of business process which can make a company. Sales is considered as a primary business process because it is the ultimate revenue generator for any and every business and without revenue business cannot survive or run. Most of the startups have a strong sales system in order to get more revenue and profit generation to run the organization.

Even the Conglomerates focus on the sales process in order to generate higher profits and gains or retain market share in the market. Without sales, any organization cannot survive and that is the reason why the money spent on the sales department is the second highest of the expenses in the entire organization. The company may or may not have any other department but it is essential to have a sales department. Be it a small organization or roadside vendors or a multinational organization, without sales, none of the businesses can survive.

2) Customer service:

Since sales is an essential and primary type of business process the next one is customer service. Once the sales orders are generated it is essential that there is a back in the team which process is the orders and provides them to the customers. Customer service also helps in managing the situation when customers required information about the product or service for assistance with the technicalities of the product. Good customer service can retain customers as well as bring back lost customers better than size department.

3) Finance department:

Types of Business Processes - 2

Once the business picks up and the money starts flowing and it is essential to manage the finances. There should be a dedicated department to manage the incoming and outgoing server the finances properly so that profits are eating expenses are reduced and all the employees are paid on time.

The finance department also looks after expense management and investment management. It is essential that expenses are kept under control and all the necessary expenses are performed so that the savings and the profits left after reducing all the expenses or much higher. The finance department also deals with daily activities like expense approvals salary disbursement and taking care of all the miscellaneous and recurring expenses.

4) Operations processes:

Supply chain management and operations on the primary skill are an essential part of the business and that is why are categorized under primary processes. Once the sale starts generating inventory management and stock management should be done in a proper way in order to supply the clients with the products or services. Supply chain management is also essential for proper management of deliveries and to receive deliveries stock management and warehouse management. Basics of operations are necessary without which business cannot survive.

5) Production:

Production or manufacturing of the product is of paramount importance for any business to run. It is crucial that the organization has a product or service to sell and that is where the production and manufacturing come into the picture. Once the product is designed, crafted and approved the initial production of how the product begins in the production or manufacturing facility.

This includes recruiting of the labors who are involved in the production facility right from manufacturing of the product till packaging and dispatching it. In case of services, the production and manufacturing and services would mean crafting of services are deciding on a standard flow of services. Once the services are standardized, they are delivered accordingly to the customers.

While this is true in case of the companies that provide tangible products, in case of the companies that provide intangible products or services they would have to have a ready process which the customers can avail. Having

a product or service to sell is the basic requirement of any organization. It does not matter if the product is on or not. The product may be manufactured by whom company or bought from other company and resold to the end customer, what matters is having something to sell.

B) Support Types of Business Processes

Types of Business Processes - 3

These are the types of business processes which are not involved in the delivery of the final product to the clients but they create a suitable environment for the functioning of primary processes.

They are not directly involved in generating value to the customer but support processes are important for the functioning of businesses. This process includes management processes, accounting process, human resources and such other processes to facilitate the smooth working of a company. Improving these processes makes a business strategy planning and fundamentally strong.

Following are few types of business process classified under support processes:

6) Accounting process:

The finance department is a basic requirement of every organization but Accounting processes are essential for the smooth and efficient performance of the finance department. Accounting department deals with the cash flow and the authenticity of the transactions in the organization.

It is the duty of the accounting department to let know when profits are reducing and losses are increasing. Accounting department ensures that all of the assets and liabilities are maintained correctly in the balance sheet. It also deals with accounts receivables and accounts payables, a collection of payments against sales, number of credit days to be allocated for the customer and such other functions.

7) Management process:

While it is essential that the top management is present in the company irrespective of its size, the middle management comes into the picture when there is an expansion of the organization.

Thus, middle management process falls under the secondary support process. Middle management is also responsible for getting the work done from the front line and reporting to the seniors about the work completion and target achievement. Management processes, like having a long chain of command is also a part of the support process. Work completion and delegation of authority becomes easier when Management process is present.

8) Human Resources:

Types of Business Processes - 4

While it is important to have a human resources department, it is not exactly crucial and organization working doesn't stop without having a Human resource department. Hence it is classified under the support process. Usually, the recruitment is usually done by the heads of the organization in a small company or startups. But as the workload and the corporate ladder increases are essential to have a dedicated Human resource department.

The department ensures smooth working of people in the organization, helps to resolve disputes, increases the communication between departments, encourages human values, helps in the career flow, is responsible also for appointing people in the organization induction of new candidates and smooth exit of the older candidates. Human resources deal with improving the candidates and the company equally and ensure mutual growth.

C) Management processes:

These types of business processes are similar to the support processes which do not add value to the end consumer. Management processes are concerned with orientation and monitoring and analyzing the day to day

business activities. These processes reason increasing the business by introducing your articles and incorporating innovation into the business.

These are usually goal-oriented processes which I am at designing and redesigning was achieving the tangible and intangible targets. They also help in maintaining the enterprise brand and ensure standing out in the market by providing their customers with added value in intangible terms. Management processes include leadership and executive decisions which are executed at the frontlines. Deciding on the targets, new product launches, expansions or closing of different departments.

Unit 2 – Introduction to business processes, analysis and process modelling

2.7. KT0206 - Business process modelling notation (BPMN)

What is BPMN?

Business Process Modeling Notation (BPMN) is a flow chart method that models the steps of a planned business process from end to end. A key to Business Process Management, it visually depicts a detailed sequence of business activities and information flows needed to complete a process.

Business Process Modeling Notation (BPMN) is a flow chart method that models the steps of a planned business process from end to end. A key to Business Process Management, it visually depicts a detailed sequence of business activities and information flows needed to complete a process

2.8. KT0207 - Difference between process and workflow

Key Differences Between Workflow and Process

It's not uncommon to hear "workflow" and "process" used interchangeably in the workplace. But did you know they're actually two different terms with unique purposes? That's right, workflow and process actually refer to distinct business terms. Let's take a look at the relationship between workflows vs. business processes.

What is a Workflow?

Let's start by defining what a workflow is. Workflow is a sequence of simple steps to reach a specific business goal. In simpler terms, a workflow automation software automates and streamlines repetitive tasks that are completed in a specific sequence every time. Typically, a workflow is created to visualize and orchestrate the connected stages of a process that can be performed in parallel or sequentially depending on specific rules or decisions.

Workflows constitute a model for how employees should be completing their tasks. That means that every person in the company follows the exact same steps to complete their tasks.

A solid, well-defined workflow allows data and tasks to flow through an organization. This enables employees to understand clearly their duties and managers identify bottlenecks and improvement opportunities. Generally, by automating workflows, you can save time and money that would otherwise be spent on repetitive, manual processes and tasks.

What is a Business Process?

A business process refers to a set of activities or tasks, often connected and automated, triggered by an event to carry out a predetermined specific operational goal. Each activity (e.g., a task), included in a process, is assigned to a team member or to an entire department.

Every organization should define its processes, analyze and measure the results to ensure that the process is meeting expectations and is getting improved.

There are three main categories of business processes:

- Operational processes that are essential and keep a business running
- Management processes that plan and control any operational process
- Supportive processes that support operational processes

Many businesses streamline processes through the use of Business Process Management (BPM) software. This tool allows businesses to solve large-scale problems. With a BPM software, they manage and automate their processes to operate smoothly, increase productivity and accuracy.

What is workflow and its process?

The definition of "workflow process" is this: It refers to a series of activities or tasks that must be completed sequentially or parallel to achieve a business outcome. In most cases, the process is linear and proceeds in a sequence determined by actions or pre-defined business rules

Workflows and Processes

The words "workflow" and "process" are often used interchangeably but actually have different meanings. Clear definitions of these terms are necessary for an organization seeking to optimize business process workflow. Find out more about what workflows, processes and procedures

are and how to unite these elements in an all-encompassing approach to business process management.

Workflow vs Process

A workflow consists of repeatable activities necessary to complete a task. A process refers to all of the elements necessary to accomplish a larger organizational goal. The general consensus is that workflows account for granular details up to small-scale objectives while processes refer to more comprehensive outcomes.

Terms such as “action,” “task” and “procedure” are used to describe parts of a workflow. Some parties also use the word “process” in this context. For the sake of this explanation, workflows consist of actions, tasks and procedures, whereas workflows contribute toward processes.

Processes are a top-down, large-scale approach toward assessing and interpreting the outcomes of workflows. This perspective is important for accountability but is not the best way to achieve regulatory compliance. While standards may be set and applied at the level of processes, the modifications necessary to meet external requirements must be made through the actions, tasks and procedures that form workflows.

Workflow vs Procedure

Stakeholders seeking to map out workflows may encounter confusion with regard to the terminology used to describe various levels of business management. Defining these component elements is essential for creating a useful explanatory breakdown of workflows for the purpose of optimization. The following definitions may be useful across industries and can be fine-tuned or modified for particular applications:

- **Actions:** The discrete activities performed by stakeholders or automated in a system
- **Tasks:** Series of related actions taken to achieve specified results or outcomes
- **Procedures:** Sequential tasks that form a distinct phase of a workflow

- **Workflows:** Series of actions, tasks and procedures that achieve a set outcome
- **Processes:** Workflows that contribute toward achieving larger goals or objectives

A workflow should account for all component procedures, tasks and actions. Breaking down each of these elements is the only way to determine the level of efficiency with which each part is currently being performed and identify opportunities for optimization.

It may not be possible for the leaders of an organization to recognize all of the factors that influence the performance of actions or tasks. Obtaining input from the stakeholders responsible for these elements is essential to making accurate maps or models of workflows. This information can also be helpful for determining whether job requirements are properly allocated for optimal performance or the satisfaction of regulatory requirements.

Further analysis of every aspect of a workflow is necessary to fully account for the actors, actions and resources involved. These are crucial considerations for increasing productivity or reducing redundancy and waste, which tend to be process-level priorities. When it comes to distinguishing a workflow diagram vs process flow diagram, a workflow model includes all of these small-scale elements whereas a process flow model takes a big-picture view of workflow outcomes.

How Workflows Connect to Processes

One of the guiding principles of business process management is the responsibility assignment matrix or RAM. The acronym RACI, which stands for “responsible, accountable, consulted and informed,” guides business process improvement in ways that pertain to the stakeholders whom the performance of actions, tasks and procedures are delegated to complete workflows.

The achievement and maintenance of workplace standards through initial training and ongoing role clarification connects the abstract elements of a workflow to the job responsibilities of specific stakeholders. These distinctions also encapsulate the difference between workflow and approval

process. These elements inform the standards that guide workflows and undergird processes.

Prev

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit Name 4

Unit Name 5

Next

2.9. KT0208 - Workflow concepts, types and differences

Workflow Definition

At the heart of every business are workflows. Whether you identify, monitor and manage them or not, they are driving your business forward. Understanding workflows, workflow management and the potential of each will help you to ensure the success of your organization. Below, you will find answers to questions such as what is workflow? What are some examples? And, how do you manage them?

How Do You Define Workflow?

A workflow is a series of steps related to processing data. It is similar to the closely related concept of a business process. However, workflows specifically focus on data and often are driven by documents and reports. Additionally, unlike processes, workflows may not be repeatable (although they often are).

Every type of organization and industry involves workflows of some type. They may be entirely human-based, mostly system-driven or somewhere in between. Anytime data is passing between two entities, it is likely a workflow.

Types of Workflows

There are a few different ways to categorize workflows. However, breaking them down by repeatability is a great way to take a deeper look at workflows.

- **Process:** A process is a repeatable series of steps used to advance business operations. In most cases, this intersects with a workflow making a process workflow. These workflows are highly predictable

and have relatively consistent data inputs and outputs. There are few variations once the workflow has begun and those can be predicted based on the inputs.

- **Case:** A case workflow is the opposite of a process workflow. It is a series of tasks needed to process data in a one-off situation. The basic workflow may be somewhat consistent between workflows. However, the input data can greatly impact the steps involved. For example, an asset inspection may be a case workflow because the follow-up steps depend on what work was needed to fix issues with the asset.
- **Project:** A project workflow is somewhere in between a case and a process in terms of repeatability. Each project is a little different. However, the data workflow is roughly the same and somewhat predictable. The follow-up steps from the data input will involve branching but can still be mapped with relative ease.

Workflow Examples

There are many workflows in the business world. Any series of steps that involves processing data is a workflow. The following are some real-world examples of workflows:

Equipment Inspection: The workflow begins with a technician opening the inspection checklist document. He or she then follows the included steps, completing an inspection report in the process. The report is submitted and forwarded to the relevant managers. This can include the inspectors' manager as well as any in charge of follow-up actions. The next steps depend on the results of the inspection. Necessary repairs and maintenance can be scheduled as required.

Performance Reporting: This workflow may be periodic or ongoing, depending on the needs of the organization. Performance data can be automatically gathered from systems used by the relevant team. That data can then be compiled into a report to be reviewed by a manager. This workflow is a repeatable process that is mostly driven by software.

Team Member Onboarding: Onboarding a new team member is a common workflow. It starts with sending forms to the new employee and the

supervisor. These are completed with relevant information about the employee and his or her employment. That information is then processed by human resources. Once ready, other teams can follow up, such as legal and IT.

Prev

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit Name 4

Unit Name 5

Next

Unit 2 – Introduction to business processes, analysis and process modelling

2.10. KT0209 - Workflow-based logic

Workflow Logic allows you to route your form forward or backward in your workflow process based on how the form is filled out and completed. For example if you have a workflow for employee reviews. If the employee is granted a pay raise by their manager the form needs to go to the HR team for approval.

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit Name 4

Unit Name 5

Next

Unit 2 – Introduction to business processes, analysis and process modelling

2.11. KT0210 - Workflow tools

What are workflow tools used for?

Workflow tools are the business user's answer to BPM (Business Process Management) software because they are a human-centric approach to managing and automating common business processes

Best Workflow Tools

- Kiss flow Workflow
- Nintex
- Process Maker
- bpm'online
- Flokzu

Unit 2 – Introduction to business processes, analysis and process modelling

2.12. KT0211 - Workflow based solutions or functionality

What is a workflow solution?

A workflow solution is a software designed to optimize business processes. It does this by standardizing everyday predictable workplace activities as 'workflows'. Workflow solutions are used to both streamline business processes as well as automate them.

What is a workflow-based system?

A workflow system provides a platform for automating processes efficiently. A workflow system is a platform that combines several discrete workflow tools into one cohesive application that automates processes involving both machine and human tasks, usually in a linear sequence.

Unit 2 – Introduction to business processes, analysis and process modelling

2.13. KT0212 - Typical major workflows to streamline and automate

What workflows can be automated?

HR can automate workflows such as time sheets, onboarding and offboarding employees or managing other changes. In general, workflow automation can also be used to schedule the uses of specific resources, monitor access to specific rooms and areas or approve invoices.

What is automation or workflow streamlining?

Workflow automation is about improving speed, accuracy, and efficiency within a business process. When you automate workflows, you reduce the level of manual work that your employees have to do, freeing them up for more value-added tasks. Automated workflow systems also reduce the potential for mistakes

Workflow automation is an approach to making the flow of tasks, documents and information across work-related activities perform independently in accordance with defined business rules. When implemented, this type of automation should be a straightforward process that is executed on a regular basis to improve everyday productivity.

Criteria for deciding when to use workflow automation include the following:

The task is repetitive.

The task needs to be achieved accurately, without any chance of human error.

A series of simple tasks can be made more efficient when automated.

Workflow automation should make it easier for an organization to streamline its workflows and identify other areas that can be automated to increase efficiency. For example, some automated workflow tasks can be managing spreadsheets or emails.

Importance of workflow automation

Workflows should be automated whenever possible for numerous reasons, including faster operations and an increase in efficiency and accuracy. Other improvements include the following:

- This emanates from increased task efficiency, allowing employees to work on other, nonautomated tasks.
- Cost savings. The savings are due to increased productivity.
- Visibility. If workflow mapping is implemented, then automation processes should be more transparent, giving an organization a top-down view of its workflows.
- Communication improvements. If visibility is increased, then communication for employees can be more accurate.
- Better customer service. This can be provided by automating responses to customer complaints, for example.
- Potential to increase customer engagement. Customers might respond quicker using automation tools.
- This can be improved because of an increased and mapped-out visibility of workflows.
- Ridding redundancies. Workflow redundancies -- any task that is unnecessary -- can be identified more readily.
- Improved overall end product. Human error is taken out of the equation.
- Digital workflow can be tracked. This allows an organization to review how well its business operates.

Benefits of workflow automation

Benefits of workflow automation include the following:

- reduced workflow cycles;
- less need for manual labor;
- less need for manual handling of products;
- more visibility;
- more visibility means an easier time identifying operational bottlenecks;
- improved customer satisfaction when focus is placed on customers;
- overall employee satisfaction, which eliminates the need for potentially dull, repetitive tasks;
- improvements in employee satisfaction via providing workflow analysis tools, which can include dashboards and key performance indicators (KPIs);
- better internal and external communications;
- more accountability for who is responsible for what in an organization resulting from each step in a business workflow being clearly assigned to one action;
- providing employees time to manage other tasks;
- increased production;
- saved costs;
- less potential for human error;
- scalability, because workflow automations can be changed and added whenever needed; and
- more efficient task management, with the inclusion of dashboards, calendars and other tools that can be made available through workflow automation software tools.

Uses of workflow automation

Workflow automation can be used in industries and departments such as in healthcare, legal, DevOps, finance, marketing, sales, IT and human resources (HR). The following shows examples of workflow automation in specific industries and areas:

For healthcare industries, workflow automation can be used in the automation of staff work schedules, as well as on-call rotations. Workflow automation can also be of aid in patient admission and discharge, as well as transferring patients' electronic health records automatically.

In legal departments, workflow automation can be used to automate billing, input new client information, submit and track contract reviews, and manage case deadlines.

In DevOps teams, automated workflows might include orchestration of the software development pipeline, monitoring and collecting data, developing testing code, and the deployment of tests and code. As an example, a DevOps group could automate tests of an e-commerce app.

Introduction to Data Science Programs

Unit 3

Unit Overview

The following topics are covered in this unit:

- The three most popular languages in data science
- Underlying design philosophy, grammar and data structures
- Data science capabilities of above programs
- Importance of programming and programming languages
- Productivity (e.g. GitHub, git, Unix/Linux, and RStudio)

Assessment Criteria

At the end of this unit the student should be able to

- Learn about the various data science programs and tools available for conducting sophisticated data analysis.

3.1. Welcome to the Introduction to Data Science Programs Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

3.1.1. Introduction to this Unit

Data science relies heavily on programming languages for data analysis, modeling, and visualization. The three most popular languages in data science—Python, R, and SQL—are essential tools for data professionals. Python stands out for its simplicity and versatility, offering a rich ecosystem of libraries like pandas, NumPy, and scikit-learn for data manipulation, machine learning, and visualization. R is renowned for its statistical analysis and graphical capabilities, making it a favorite among statisticians. SQL, meanwhile, is critical for querying and managing structured data in databases, forming the backbone of many data workflows.

Each of these languages has a unique design philosophy and grammar that caters to specific data science tasks. Python emphasizes readability and flexibility, with straightforward syntax and support for diverse data structures like lists, dictionaries, and arrays. R is built around statistical computing and visualization, providing robust tools for handling data frames and performing advanced analyses. SQL uses a declarative approach to interact with relational databases, making it indispensable for data extraction and manipulation. Together, these languages enable comprehensive data science capabilities, from exploratory data analysis to predictive modeling and data storytelling.

Programming proficiency is central to data science, as it underpins the ability to clean, analyze, and interpret data effectively. Tools that enhance productivity, such as GitHub, git, Unix/Linux, and RStudio, are essential for collaborative development and efficient workflows. Git and GitHub enable version control and team collaboration, while Unix/Linux offers powerful command-line utilities for managing data and scripts. RStudio, tailored for R, provides an integrated development environment (IDE) that streamlines coding, visualization, and reporting. Mastering these tools and languages equips data scientists with the skills to tackle complex problems and contribute meaningfully to data-driven projects.

3.2. KT0301 - The three most popular languages in data science

Top data science programming languages

- Python
- R
- SQL
- Java
- Julia
- Scala C/C++
- JavaScript
- Swift
- Go
- MATLAB
- SAS

3.3. KT0302 - Underlying design philosophy, grammar and data structures

What is grammar in data structure?

It is a finite set of formal rules for generating syntactically correct sentences or meaningful correct sentences. Constitute Of Grammar : Grammar is basically composed of two basic elements – Terminal Symbols

Grammar:

It is a finite set of formal rules for generating syntactically correct sentences or meaningful correct sentences.

Constitute Of Grammar:

Grammar is basically composed of two basic elements –

Terminal Symbols –

Terminal symbols are those which are the components of the sentences generated using a grammar and are represented using small case letter like a, b, c etc.

Non-Terminal Symbols –

Non-Terminal Symbols are those symbols which take part in the generation of the sentence but are not the component of the sentence. Non-Terminal Symbols are also called Auxiliary Symbols and Variables. These symbols are represented using a capital letter like A, B, C, etc.

Formal Definition of Grammar :

Any Grammar can be represented by 4 tuples – $\langle N, T, P, S \rangle$

N – Finite Non-Empty Set of Non-Terminal Symbols.

T – Finite Set of Terminal Symbols.

P – Finite Non-Empty Set of Production Rules.

S – Start Symbol (Symbol from where we start producing our sentences or strings).

Production Rules :

A production or production rule in computer science is a rewrite rule specifying a symbol substitution that can be recursively performed to generate new symbol sequences. It is of the form $\alpha \rightarrow \beta$ where α is a Non-Terminal Symbol which can be replaced by β which is a string of Terminal Symbols or Non-Terminal Symbols.

Example-1 :

Consider Grammar $G1 = \langle N, T, P, S \rangle$

$T = \{a, b\}$ #Set of terminal symbols

$P = \{A \rightarrow Aa, A \rightarrow Ab, A \rightarrow a, A \rightarrow b, A \rightarrow \epsilon\}$ #Set of all production rules

$S = \{A\}$ #Start Symbol

3.4. KT0303 - Importance of programming and programming languages

Programming languages use classes and functions that control commands. The reason that programming is so important is that it directs a computer to complete these commands over and over again, so people do not have to do the task repeatedly. Instead, the software can do it automatically and accurately.

What Is Programming?

Programming is using a language that a machine can understand in order to get it to perform various tasks. Computer programming is how we communicate with machines in a way that makes them function how we need.

What Is a Program?

A program is a group of logical, mathematical and sequential functions grouped together. When they are grouped, these functions perform a task. Each programming language focuses on different types of tasks as well as gives commands to the machine in different ways.

What Is a Class?

In computer programming, a class contains a group of instructions that act as commands for the computer. The class is made up of variables, integers, decimals and other symbols. These are put together in certain orders to let the computer know what task to perform.

What Is a Function?

Even if you are new to computer programming, you are familiar with functions. If you use an online music streaming program, you press the button to start or pause the play. Those are functions.

When classes of programming languages are grouped together, they create functions. These functions allow you to perform certain tasks in a program. Some functions are small and control just one aspect of a piece of software or program. Other functions are big and ensure that the program itself runs.

What Is a Command?

Commands are the methods to control certain aspects of the program or machine. Programming languages use classes and functions that control commands. The reason that programming is so important is that it directs a computer to complete these commands over and over again, so people do not have to do the task repeatedly. Instead, the software can do it automatically and accurately.

Why Is it Important to Know About Computer Programming?

If you are thinking about earning your computer programming degree, you will need to know about programming languages, classes, functions and commands. You will create applications, software or different programs. In addition, you may create programs that need to work on various operating systems such as iOS or Android. Those programs have different functions and classes, which means they rely on different programming languages.

All applications on the web are created using computer programming. The languages used in each application you have range from similar to vastly different. Additionally, some languages create things that are running in the background, so you do not even know they are there. Learning computer programming languages allows you to be a versatile computer programmer.

What Is the Future Impact of Computer Programming?

Technology production is an essential part of an evolving world. This means that computer programming is exceptionally important for our future as a global society. Computer programming degree graduates can help create this future by automating processes, collecting data, analyzing information and sharing knowledge to continuously innovate and improve upon existing processes.

This means that, while computer programming is extremely important today, it may be even more impactful in the future. As computer programmers across the world work to learn new ways of communicating with machines and computers, the field will continue to grow. Earning your computer programming degree now means you can be part of that research and testing to develop functions that can help society.

Computer programming is important today because so much of our world is automated. Humans need to be able to control the interaction between people and machines. Since computers and machines are able to do things so efficiently and accurately, we use computer programming to harness that computing power.

What Are the Important Computer Programming Languages to Learn?

Computer programming is evolving and so are the languages that are used to develop software and applications. Different programming languages are used for different categories of developers. Some languages are best for beginners, while others are more suited for advanced computer programmers. In addition, some languages are best for different use cases such as web apps, mobile apps and distributed systems.

The best way to determine what programming languages to learn is to know what skills you will need in order to be a successful computer programmer.

Popular and important computer programming languages based on necessity and application include:

- Python
- Java
- C/C++
- JavaScript
- Swift

Each of these language ranges in usability and ease of learning. Python is considered the best beginner programming language. It is easy to learn and to deploy. Java has been a popular language for many decades. It is the official language for Android apps. C and C++ are considered the foundational languages for many operating systems and file systems. JavaScript is popular for front end developers because it helps make applications look clean and clear for the user. Swift is a native iOS language and has been growing in popularity as Mac and Apple products become industry favorites.

3.5. KT0304 - Productivity (e.g. GitHub, git, Unix/Linux, and RStudio)

How do I link my GitHub and RStudio?

Test Drive RStudio and GitHub

- **Step 1:** Make a new repo on GitHub. Go to GitHub.com and login. ...
- **Step 2:** Clone the new GitHub repository to your computer via RStudio. ...
- **Step 3** plan B: Connect a local RStudio project to a GitHub repo. ...
- **Step 4:** Make local changes, save, commit. ...
- **Step 5:** Push your local changes online to GitHub.

Unit Overview

The following topics are covered in this unit:

- Data types and variables
- Operators
- Conditional statements
- Loops
- Script
- Functions
- Probability
- Inference and modelling

Learning Outcomes

At the end of this unit the student should be able to

- Dive deep into data analytics techniques, including statistical analysis, predictive modeling, and exploratory data analysis.

4.1. Welcome to the Data Analytics Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

4.1.1. Introduction to this Unit

Data analytics is a cornerstone of data-driven decision-making, involving the systematic analysis of datasets to uncover insights and patterns. It starts with understanding data types and variables, which include categorical, numerical, and ordinal data, as well as dependent and independent variables. Operators, such as arithmetic, comparison, and logical operators, are fundamental for performing calculations and evaluating conditions within datasets. Conditional statements like if-else allow analysts to implement decision-making logic, while loops such as for and while automate repetitive tasks, enhancing efficiency in data handling and analysis.

Scripts and functions are essential tools for organizing and reusing code in data analytics workflows. Scripts serve as containers for executing sequences of instructions, making complex tasks repeatable and modular. Functions, on the other hand, encapsulate specific operations into reusable blocks, improving readability and reducing redundancy in code. Together, these programming constructs enable data analysts to build scalable and efficient solutions for data manipulation and analysis.

A foundational element of data analytics is probability, which quantifies the likelihood of events and underpins statistical inference and modeling. By applying concepts like probability distributions, sampling, and hypothesis testing, analysts can draw meaningful conclusions from data. Modeling techniques, such as regression and classification, further allow for the prediction of outcomes and the identification of relationships between variables. These techniques form the basis of advanced analytics, empowering organizations to make data-informed decisions and solve real-world problems effectively.

Unit 4 – Data Analytics

4.2. KT0401 - Data types and variables

What a data analytics do?

Data analytics helps individuals and organizations make sense of data. Data analysts typically analyse raw data for insights and trends. They use various tools and techniques to help organizations make decisions and succeed.

A variable can be thought of as a memory location that can hold values of a specific type. The value in a variable may change during the life of the program—hence the name “variable.”

What are the 5 data analytics?

The Five Key Types of Big Data Analytics Every Business Analyst Should Know

- Prescriptive Analytics.
- Diagnostic Analytics.
- Descriptive Analytics.
- Predictive Analytics.
- Cyber Analytics.
 - Interested in learning more about business analytics and data science?

4.3. KT0402 - Operators

What are operators in data analytics?

An operator is a character or set of characters that can be used to perform the desired operation on the operands and produce the final result. The final result completely depends on the type of operators used. For example, consider you want to perform a mathematical calculation of $5+3$.

4.4. KT0403 - Conditional statements

What is conditional statement in data structure?

A conditional statement is used to determine whether a certain condition exists before code is executed. Conditional statements can help improve the efficiency of your code by providing you with the ability to control the flow of your code, such as when or how code is executed.

What are conditional statements in programming?

Conditional statements are used through the various programming languages to instruct the computer on the decision to make when given some conditions. These decisions are made if and only if the pre-stated conditions are either true or false , depending on the functions the programmer has in mind.

4.5. KT0404 - Loops

What is for loop in data science?

A for loop executes commands once for each value in a collection. Doing calculations on the values in a list one by one is as painful as working with pressure_001 , pressure_002 , etc. A for loop tells Python to execute some statements once for each value in a list, a character string, or some other collection

Data Loop is a continuous, iterative process of capturing and analyzing data, getting valuable insights, translating into action items and 'looping' the process all over again.

4.6. KT0405 - Script

What is script called?

A scripting language can be viewed as a domain-specific language for a particular environment; in the case of scripting an application, it is also known as an extension language.

Do you mean by script?

A script is a written version of a play or movie. If you're auditioning for a movie, you'll get the script to practice a scene or two. Script comes from the Latin *scrībĕre*, meaning "to write," and all its meanings have to do with something written. Your handwriting is your script.

4.7. KT0406 - Functions

What is data analytics function?

Functions allow you to automate common tasks in a more powerful and general way than copy-and-pasting. Writing a function has three big advantages over using copy-and-paste: You can give a function an evocative name that makes your code easier to understand.

What are the functions of data?

Data functions are the nuts and bolts of the digital world. In development, they allow programmers to exchange directly with a database and modify object in-line. In Excel, they query, create, and modify arrays. In math, they allow statisticians to address datasets using Sigma (Σ) and Pi (Π) notation.

4.8. KT0407 - Probability

Probability is simply how likely something is to happen. Whenever we're unsure about the outcome of an event, we can talk about the probabilities of certain outcomes—how likely they are. The analysis of events governed by probability is called statistics.

4.9. KT0408 - Inference and modelling

What is inference and Modelling?

Models are constructed using accepted theoretical principles, prior knowledge and expert judgement. Inference is the process by which we compare the models to the data. This normally involves casting the model mathematically and using the principles of probability to quantify the quality of match.

Unit Overview

The following topics are covered in this unit:

- Importing data from different file formats
- Web scraping
- How to tidy data using suitable software packages to better facilitate analysis
- String processing with regular expressions (regex)
- HTML parsing
- Wrangling data using suitable software package
- How to work with dates and times as file formats
- Text mining

Learning Outcomes

At the end of this unit the student should be able to

- Master the art of data wrangling, preparing and cleaning data to make it suitable for analysis.

5.1. Welcome to the Wrangling Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- **Review:** Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- **Study your prescribed material**
- **Follow the Study Material References:** Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

5.1.1. Introduction to this Unit

Data wrangling is the process of transforming and organizing raw data into a structured format suitable for analysis. A crucial step in this process is importing data from different file formats, such as CSV, Excel, JSON, and databases, using tools like Python's pandas or R's readr package. For data sourced from websites, web scraping techniques enable the extraction of information using libraries such as BeautifulSoup in Python. This is often paired with HTML parsing to interpret and navigate HTML documents, facilitating the collection of structured data from web pages.

Once data is imported, tidying data becomes essential to enhance its usability. This involves reshaping datasets into a standardized format, such as ensuring each column represents a variable and each row represents an observation. Tools like Python's pandas or R's tidyr package simplify this process. Additionally, string processing with regular expressions (regex) plays a key role in cleaning textual data, such as extracting patterns, removing unwanted characters, or standardizing formats.

Advanced wrangling tasks include working with dates and times, where software packages like Python's datetime module or R's lubridate facilitate handling complex time formats. Similarly, text mining enables the extraction of insights from unstructured text, such as sentiment analysis or keyword extraction. Wrangling tools like OpenRefine or scripting in Python or R streamline these tasks, ensuring data is clean, consistent, and ready for analysis. By mastering these techniques, analysts can transform chaotic datasets into valuable assets for decision-making.

Unit 5 – Wrangling

5.2. KT0501 - Importing data from different file formats

File formats supported for importing data

Koordinates support multiple file formats for both import and export. Below, we provide a brief introduction to some of the file formats we support.

A note on importing data

Note that, when importing data, you are able to put multiple types of data into a single data archive. Koordinates will extract them and either create individual data layers or merge them into a single layer.

Quick overview of import file formats

- Vector and tabular
- Shapefile (.SHP)
- MapInfo TAB (.TAB)
- Esri File Geodatabase
- GeoPackage / SQLite
- GeoJSON
- CSV
- XLS, XLSX (Excel Spreadsheet) and ODS (OpenDocument Spreadsheet)
- Esri Coverage

Image/Raster

- GeoJPEG
- GeoTIFF, TIFF with world files (.TFW)
- JPEG2000
- Gridded Raster
- ESRI Binary Grid
- ESRI ASCII Grid

- GeoTIFF

If your data is in a format not listed above, you may be able to convert it into a supported format.

MapInfo TAB

MapInfo TAB is a proprietary format used for vector datasets, and is primarily for use in the suite of MapInfo GIS applications.

Like other GIS file formats, TAB files contain a few different kinds of file formats, which each do a different work. These include:

- .tab: The main file that links all the other files in the dataset
- .dat: This file stores the attribute data - that is, the the non-spatial data in your dataset
- .id: This links graphic data to the database information
- .map: This has the geographic information that enables the data to be represented on a map.
- .ind: This is an index file for tabular data

GeoTIFF

A GeoTIFF file is a georeferenced TIFF, which means it can have geographic information such as map projections embedded within the TIFF itself. This means that a GIS application can position the image in the correct location

TIFF is an open and non-proprietary file format that is widely used for raster imagery and aerial photography. Strictly speaking, GeoTIFF is a metadata format, but the TIFF format enables both the data and metadata to exist in the same file.

Shapefile (SHP)

Shapefile is a GIS format developed for vector datasets. It should work in most common GIS applications.

Your Shapefile download will contain a range of different file formats, which each do a different kind of work within your application. These are:

- .shp: this file represents the feature geometry - i.e. the points, lines and polygons in spatial dataset.
- .shx: this file represents the 'shape index position,' and is used to search forward and backwards.
- .dbf: this is the database file, which contains attribute data and object IDs.
- .prj: this file will contain the projection information.
- .cpg: this identifies the character encoding for the .dbf database file.

GeoPackage / SQLite

Geopackage is a SQLite database format from the Open Geospatial Consortium, which is intended to be a modern alternative to older formats like Shapefile.

Geopackage is an open and platform independent format, and is supported in a range of applications. It is an especially useful format for development of mobile applications for smartphones and tablets.

File Geodatabase (FGDB)

File Geodatabase is a proprietary Esri database format that is used for more complex uses of GIS datasets in Esri software. This format is often used because it allows for much larger file formats and can provide better performance than Shapefiles.

GeoJPEG

GeoJPEG's are JPEG files that have geographic information attached within the file -- that is to say, it is a georeferenced JPEG. When downloaded, the

jpeg file will be accompanied by a .jgw file that will carry the associated geographic information.

JPEG 2000

JPEG 2000 is the newer version of the JPEG file format, which is intended to provide better compression and image quality than an ordinary JPEG. While JPEG 2000 has some benefits to JPEG, it is not as widely supported as JPEG.

ASCII Grid

ASCII Grid is a proprietary file format from Esri.

CSV and GeoCSV

Comma Separated Value format - or CSV - is a format that enables tabular data to be easily exchanged by different applications. GeoCSV is its geospatial extension.

5.3. KT0502 - Web scraping

What is web scraping?

Web scraping is the process of collecting structured web data in an automated fashion. It's also called web data extraction. Some of the main use cases of web scraping include price monitoring, price intelligence, news monitoring, lead generation, and market research among many others.

In general, web data extraction is used by people and businesses who want to make use of the vast amount of publicly available web data to make smarter decisions.

If you've ever copied and pasted information from a website, you've performed the same function as any web scraper, only on a microscopic, manual scale. Unlike the mundane, mind-numbing process of manually extracting data, web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet's seemingly endless frontier.

How do you use a data scraper?

Whether you're using a data scraper tool yourself or outsourcing the job to a web data extraction specialist, you'll need to know a bit more about the differences between web crawling and web scraping. Just as importantly, you'll need to understand the possible pitfalls of extraction and how to avoid them. Read on to find out how web scraping works and how to achieve it successfully.

Web scraping is popular

And it should not be surprising because web scraping provides something really valuable that nothing else can: it gives you structured web data from any public website.

More than a modern convenience, the true power of data web scraping lies in its ability to build and power some of the world's most revolutionary business applications. 'Transformative' doesn't even begin to describe the way some companies use web scraped data to enhance their operations, informing executive decisions all the way down to individual customer service experiences.

What is data scraping good for?

Web data extraction – also widely known as data scraping – has a huge range of applications. A data scraping tool can help you automate the process of extracting information from other websites, quickly and accurately. It can also make sure the data you've extracted is neatly organized, making it easier to analyze and use for other projects.

In the world of e-commerce, web data scraping is widely used for competitor price monitoring. It's the only practical way for brands to check the pricing of their competitors' products and services, allowing them to fine-tune their own price strategies and stay ahead of the game. It's also used as a tool for manufacturers to ensure retailers are compliant with pricing guidelines for their products. Market research organizations and analysts depend on web data extraction to gauge consumer sentiment by keeping track of online product reviews, news articles, and feedback.

There's a vast array of applications for data extraction in the financial world. Data scraping tools are used to extract insight from news stories, using this information to guide investment strategies. Similarly, researchers and analysts depend on data extraction to assess the financial health of companies. Insurance and financial services companies can mine a rich seam of alternative data scraped from the web to design new products and policies for their customers.

Applications for web data extraction don't end there. Data scraping tools are widely used in news and reputation monitoring, journalism, SEO monitoring, competitor analysis, data-driven marketing and lead

generation, risk management, real estate, academic research, and much more.

The basics of web scraping

It's extremely simple, in truth, and works by way of two parts: a web crawler and a web scraper. The web crawler is the horse, and the scraper is the chariot. The crawler leads the scraper, as if by hand, through the internet, where it extracts the data requested. Learn the difference between web crawling & web scraping and how they work.

The crawler

A web crawler, which we generally call a "spider," is an artificial intelligence that browses the internet to index and search for content by following links and exploring, like a person with too much time on their hands. In many projects, you first "crawl" the web or one specific website to discover URLs which then you pass on to your scraper.

The scraper

A web scraper is a specialized tool designed to accurately and quickly extract data from a web page. Web scrapers vary widely in design and complexity, depending on the project. An important part of every scraper is the data locators (or selectors) that are used to find the data that you want to extract from the HTML file - usually, XPath, CSS selectors, regex, or a combination of them is applied.

What is a web scraping tool?

A web scraping tool is a software program that's designed specifically to extract (or 'scrape') relevant information from websites. You'll almost certainly be using some kind of scrape tool whenever you are collecting data from web pages programmatically.

A scraping tool typically makes HTTP requests to a target website and extracts the data from a page. Usually, it parses content that is publicly accessible and visible to users and rendered by the server as HTML. Sometimes it also makes requests to internal application programming interfaces (APIs) for some associated data – like product prices or contact details – that are stored in a database and delivered to a browser via HTTP requests.

There are various kinds of web scrape tools out there, with capabilities that can be customized to suit different extraction projects. For example, you might need a scraping tool that can recognize unique HTML site structures, or extract, reformat and store data from APIs.

Scraping tools can be large frameworks designed for all kinds of typical scraping tasks, but you can also use general-purpose programming libraries and combine them to create a scraper.

For example, you might use an HTTP requests library - such as the Python-Requests library - and combine it with the Python BeautifulSoup library to scrape data from your page. Or you may use a dedicated framework that combines an HTTP client with an HTML parsing library. One popular example is Scrapy, an open-source library created for advanced scraping needs.

Unit 5 – Wrangling

5.4. KT0503 - How to tidy data using suitable software packages to better facilitate analysis

What makes a data set tidy?

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

What is tidy data in Python?

Tidy Data is a way of structuring datasets to facilitate analysis. In 2014, Hadley Wickham published an awesome paper named Tidy Data, that describes the process of tidying a dataset in R. My goal with this article is to summarize these steps and show the code in Python.

Unit 5 – Wrangling

5.5. KT0504 - String processing with regular expressions (regex)

What is the regular expression for string?

A regular expression (shortened as regex or regexp; also referred to as rational expression) is a sequence of characters that specifies a search pattern in text. Usually such patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation.

A regular expression (shortened as regex or regexp; also referred to as rational expression) is a sequence of characters that specifies a search pattern in text. Usually such patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation. It is a technique developed in theoretical computer science and formal language theory.

5.6. KT0505 - HTML parsing

Parsing means analyzing and converting a program into an internal format that a runtime environment can actually run, for example the JavaScript engine inside browsers. The browser parses HTML into a DOM tree. HTML parsing involves tokenization and tree construction.

A computer program that parses content is called a parser. There are in general 2 kinds of parsers

- **Top-down parsing**- Top-down parsing can be viewed as an attempt to find left-most derivations of an input-stream by searching for parse trees using a top-down expansion of the given formal grammar rules. Tokens are consumed from left to right. Inclusive choice is used to accommodate ambiguity by expanding all alternative right-hand-sides of grammar rules.
- **Bottom-up parsing** - A parser can start with the input and attempt to rewrite it to the start symbol. Intuitively, the parser attempts to locate the most basic elements, then the elements containing these, and so on. LR parsers are examples of bottom-up parsers. Another term used for this type of parser is Shift-Reduce parsing.

Unit 5 – Wrangling

5.7. KT0506 - Wrangling data using suitable software package

What Is Data Wrangling?

Data wrangling is the act of cleaning, structuring, and transforming raw data into formats that simplify data analytics processes. Data wrangling often involves working with messy and complex data sets that are not ready for data pipeline processes. Data wrangling moves raw data to a refined state or refined data to optimized state and production-ready level.

Some of the known tasks in data wrangling include:

- Merging multiple datasets into one large dataset for analysis.
- Examining missing/gaps in data.
- Removing outliers or anomalies in datasets.
- Standardizing inputs.

The large data stores involved in data wrangling processes are usually beyond manual tuning, necessitating automated data preparation methods to produce more accurate and quality data.

Goals of Data Wrangling

- Besides preparing data for analysis as the bigger goal, other goals include:
- Creating valid and novel data out of messy data to drive decision-making in businesses.
- Standardizing raw data into formats that Big Data systems can ingest.
- Reducing the time spent by data analysts when creating data models by presenting orderly data.
- Creating consistency, completeness, usability, and security for any dataset consumed or stored in a data warehouse.

Common approaches to Data Wrangling

- **Discovering**

Before data engineers start data preparation tasks, they need to understand how it is stored, the size, what records are kept, the encoding formats, and other attributes describing any dataset.

- **Structuring**

This process involves organizing data to take readily usable formats. Raw datasets may need structuring in how the columns appear, the number of rows, and tuning other data attributes to simplify analysis.

- **Cleaning**

Structured datasets need to be gotten rid of inherent errors and anything that can skew the data within. Cleaning thus entails removing multiple cell entries with similar data, deleting empty cells and outlier data, standardizing inputs, renaming confusing attributes,

- **Enriching**

Once data has passed the structuring and cleaning stages, it is necessary to assess data utility and augment it with values from other datasets lacking to give the desired data quality.

- **Validating**

The validating process entails iterative programming aspects that shed light on data quality, consistency, usability, and security. Validating phase ensures all transformation tasks are achieved and flags datasets as ready for analytics and modeling phases.

- Presenting

After all the stages are passed, the wrangled datasets are presented/shared within an organization for analytics. Documentation of preparation steps and metadata generated along the wrangling process is also shared in this stage.

Unit 5 – Wrangling

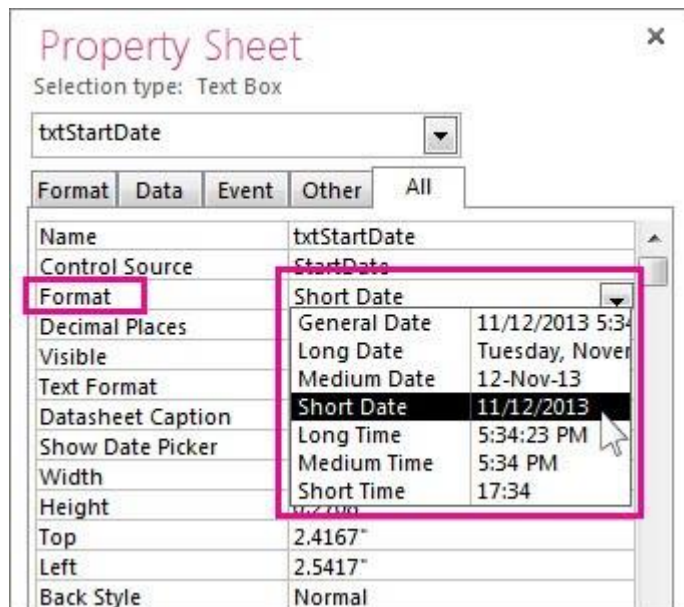
5.8. KT0507 - How to work with dates and times as file formats

Apply a predefined format

Access provides several predefined formats for date and time data.

In a table

1. Open the table in Design View.
2. In the upper section of the design grid, select the Date/Time field that you want to format.
3. In the Field Properties section, click the arrow in the Format property box, and select a format from the drop-down list.



4. After you select a format, the Property Update Options button appears, and lets you to apply your new format to any other table fields and form controls that would logically inherit it. To apply your changes throughout the database, click the smart tag, and then click Update Format everywhere <Field Name> is used. In this case, Field Name is the name of your Date/Time field.
5. To apply your changes to the entire database, when the Update Properties dialog box appears and displays the forms and other objects that will inherit the new format. Click Yes.
6. Save your changes and switch to Datasheet view to see whether the format meets your needs.

Note New forms, reports, or views that you create based on this table get the table's formatting, but you can override this on the form, report, or view without changing the table's formatting.

In a form or report

1. Open the form or report Layout View or Design View.
2. Position the pointer in the text box with the date and time.
3. Press F4 to display the Property Sheet.
4. Set the Format property to one of the predefined date formats.

In a query

1. Open the query in Design View.
2. Right-click the date field, and then click Properties.
3. In the Property Sheet, select the format you want from the Format property list.

In an expression

- Use the Format DateTime function to format a date value into one of several predefined formats.

You might find this helpful if you are working in an area that requires an expression, such as a macro or a query.

Prev

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit Name 4

Unit Name 5

Next

5.9. KT0508 - Text mining

What is text mining?

Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. By applying advanced analytical techniques, such as Naïve Bayes, Support Vector Machines (SVM), and other deep learning algorithms, companies are able to explore and discover hidden relationships within their unstructured data.

Text is a one of the most common data types within databases. Depending on the database, this data can be organized as:

- **Structured data:** This data is standardized into a tabular format with numerous rows and columns, making it easier to store and process for analysis and machine learning algorithms. Structured data can include inputs such as names, addresses, and phone numbers.
- **Unstructured data:** This data does not have a predefined data format. It can include text from sources, like social media or product reviews, or rich media formats like, video and audio files.
- **Semi-structured data:** As the name suggests, this data is a blend between structured and unstructured data formats. While it has some organization, it doesn't have enough structure to meet the requirements of a relational database. Examples of semi-structured data include XML, JSON and HTML files.

Unit Overview

The following topics are covered in this unit:

- Introduction to data structures
- Identifying data structures
- Assigning values to data structures
- Data manipulation

Learning Outcomes

At the end of this unit the student should be able to

- Understand the different data structures used in data analysis and how they influence the analysis process.

6.1. Welcome to the Data Structures Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

6.1.1. Introduction to this Unit

Data structures are foundational components in programming and data science, designed to organize, store, and manage data efficiently. They provide the means to handle and manipulate data in various formats, ensuring that operations such as searching, sorting, and updating are optimized. Common data structures include arrays, lists, stacks, queues, hash tables, trees, and graphs. Each structure serves a specific purpose, allowing developers to choose the most suitable option based on the problem at hand.

Identifying and assigning values to data structures involves selecting the appropriate structure for the data's nature and requirements. For example, lists or arrays are ideal for ordered collections, while dictionaries or hash tables are suited for key-value mappings. Assigning values is done by defining the data structure and populating it with data, such as adding elements to a list or inserting key-value pairs into a dictionary. Understanding the properties of each data structure—such as mutability, ordering, and indexing—guides their effective use.

Data manipulation encompasses operations performed on data structures to transform or analyze their contents. This includes tasks such as appending, removing, or updating elements, as well as sorting, filtering, or merging datasets. Libraries like Python's pandas or NumPy, and functions in R, provide powerful tools for advanced manipulation of data stored in structured formats. By mastering data structures and their manipulation, programmers and data analysts can ensure efficiency, accuracy, and scalability in their workflows.

6.2. KT0601 - Introduction to data structures

Data Structures are a specialized means of organizing and storing data in computers in such a way that we can perform operations on the stored data more efficiently. Data structures have a wide and diverse scope of usage across the fields of Computer Science and Software Engineering.

6.3. KT0602 - Identifying data structures

What are the 5 key data structures?

Data Structures

- Linear: arrays, lists.
- Tree: binary, heaps, space partitioning etc.
- Hash: distributed hash table, hash tree etc.
- Graphs: decision, directed, acyclic etc.

6.4. KT0603 - Assigning values to data structures

How do you assign a value to a structure?

How to assign values to structure members?

- Using Dot(.) operator `var_name.memeber_name = value;`
- All members assigned in one statement `struct struct_name var_name = {value for memeber1, value for memeber2 ... so on for all the members}`
- Designated initializers – We will discuss this later at the end of this post.

6.5. KT0604 - Data manipulation

What is Data Manipulation?

Data manipulation is the process of changing or altering data in order to make it more readable and organized. For example, you can arrange data alphabetically to expedite the process of finding useful information. Another example of data manipulation is website management. Website owners can use web server logs to locate the most viewed web pages, traffic sources, and much more. Similarly, stockbrokers use data manipulation to forecast stock market trends.

Why Use Data Manipulation?

Businesses use data for predicting trends, understanding customer behavior, increasing productivity, reducing costs, etc. through manipulation of data. Other additional benefits of data manipulation include:

- **Format consistency:** Data organized in a unified, orderly manner help business users make better decisions.
- **Historical overview:** Accessing data of previous projects quickly can help an organization make decisions regarding deadline projection, team productivity, budget allocation, etc.
- **Improved efficiency:** By having more organized data, a business can isolate and even reduce external variables to contribute to the overall efficiency of the business.

Data Manipulation Language

It is possible to make data more organized or readable through DML or data manipulation language. DML is a computer programming language that is used for inserting, omitting, and altering data in a database. It makes the data easy to cleanse and map for further analysis. A commonly used

data manipulation language is Structured Query Language (SQL). SQL is used to update and retrieve data in a relational database.

Manipulate Data Using Built-in Transformations

Simplify data manipulation through drag-and-drop mappings and pre-built transformations

What are Data Manipulation Tools?

Data manipulation tools allow you to modify data to make it easier to read or organize. These tools help identify patterns in your data that may otherwise not be obvious. For instance, you can arrange a data log in alphabetical order using a data manipulation tool so that discrete entries are easier to find.

People often confuse data manipulation with ETL and other transformation techniques. However, data manipulation involves sorting, rearranging, and moving data without essentially changing it. It includes operations to adapt data in the form that is needed to display information or feed and train an analytics model.

The key purpose of data manipulation is to alter the relationship (either logical or physical) that one data item has with another, not the data itself. Common operations used for data manipulation include row and column filtering, aggregation, join and concatenation, string manipulation, classification, regression, and mathematical formulas.

ETL, on the other hand, serves a different purpose. It involves extracting data from the source system and making it compatible with the destination system before writing into it.

Why Do You Need Data Manipulation Tools?

Data manipulation is a critical task in process optimization. It transforms data into a usable form that can be used further to generate insights, such as analyzing financial data, customer behavior and carrying out trend analysis.

Data manipulation is widely used during integration to make data compatible with the target system. For example, users associated with accounting often manipulate raw data acquired from vendors and marketing to comprehend product prices, sales trends, or prospective tax requirements. Similarly, stock market experts can leverage datasets to forecast market trends allowing them to manage their investment portfolios accordingly.

These are just a few use-cases of data manipulation. Some of the other ways in which manipulation can be beneficial for organizations include:

- **Data Consistency**

A consistent data format makes it easier to organize, read, and analyze data. When data comes from disparate sources, the user must transform and manipulate it to create a unified format. After standardizing the format, it is easier to write data into the enterprise system or utilize it for reporting.

- **Data Projection**

As a business, you can't deny the importance of data when it comes to business intelligence (BI). Creating an exhaustive data analysis is vital for companies, particularly when it comes to investments. Every business makes use of past data to plan for the future. Data manipulation makes it easier to create projections, especially in the financial sector, where you depend on your past investments' results for future considerations.

- **Value Generation**

Data manipulation allows you to update, modify, delete, and input data into a database. This means that you can leverage data to obtain in-depth insights and make better business decisions.

- **Redundant Data Removal**

Often, data coming from source systems include redundant, erroneous, or unwanted information. Making this data useful requires running it through

quality checks and applying cleansing filters to extract the information essential for your company. By using data manipulation, you can swiftly clean your data so that you can filter out the records that matter.

- **Data Interpretation**

When dealing with complex data that involves multiple formats and business conditions, it is next to impossible to make sense of it without manipulation. You need to have the capability to visualize data and alter it into valuable and comprehensible information. A data manipulation tool might resolve this problem by converting data into the desired format and integrate it with different tools to enhance the visual experience. This makes it easier for users to comprehend and consume data.

Tips and Steps for Data Manipulation

The most efficient way to manipulate data is via tools that offer built-in, automated data manipulation functions, such as data cleaning, mapping, aggregating, or storing. These tools save you the trouble of entering data manually and performing low-value repetitive tasks. Moreover, the automation features supported by these tools support report generation and delivery without any human intervention.

Unit Overview

The following topics are covered in this unit:

- Introduction to data visualization
- Data visualization using graphics
- Data visualization using system for declaratively creating graphics
- File formats of graphic outputs

Learning Outcomes

At the end of this unit the student should be able to

- Learn the principles of data visualization and how to use visual elements to communicate data insights effectively.

7.1. Welcome to the Data Visualization Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

7.1.1. Introduction to this Unit

Data visualization is a powerful tool for translating complex data into clear and actionable insights through visual representations such as graphs, charts, and maps. It bridges the gap between raw data and understanding, enabling audiences to quickly grasp patterns, trends, and anomalies. Effective visualization not only aids in data interpretation but also enhances communication, making it an essential skill for data analysts, scientists, and decision-makers.

Data visualization using graphics involves leveraging tools and libraries to create visual representations of data. Libraries like Matplotlib, Seaborn, and

ggplot2 allow users to craft diverse visualizations such as bar charts, scatter plots, and heatmaps. These tools provide flexibility to customize elements like colors, labels, and axes, ensuring that the graphics align with the audience's needs and the data's story.

Declarative systems like Plotly or Vega-Lite enable users to create visualizations by specifying the desired outcome rather than focusing on implementation details. These systems are particularly useful for crafting interactive and dynamic visualizations, making them ideal for dashboards and exploratory data analysis. Additionally, the file formats of graphic outputs—such as PNG, JPEG, PDF, and SVG—play a crucial role in determining the usability and scalability of visualizations. While raster formats like PNG and JPEG are suitable for static and web-based visuals, vector formats like SVG and PDF are preferred for high-resolution printing and scalability. By mastering these tools and concepts, professionals can create impactful visualizations that drive data-driven storytelling.

Unit 7 – Data Visualization

7.2. KT0701 - Introduction to data visualization

What is data visualization?

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Data visualization can be utilized for a variety of purposes, and it's important to note that is not only reserved for use by data teams. Management also leverages it to convey organizational structure and hierarchy while data analysts and data scientists use it to discover and explain patterns and trends. Harvard Business Review (link resides outside IBM) categorizes data visualization into four key purposes: idea generation, idea illustration, visual discovery, and everyday dataviz. We'll delve deeper into these below:

Data visualization is commonly used to spur idea generation across teams. They are frequently leveraged during brainstorming or Design Thinking sessions at the start of a project by supporting the collection of different perspectives and highlighting the common concerns of the collective. While these visualizations are usually unpolished and unrefined, they help set the foundation within the project to ensure that the team is aligned on the problem that they're looking to address for key stakeholders.

Data visualization for idea illustration assists in conveying an idea, such as a tactic or process. It is commonly used in learning settings, such as tutorials, certification courses, centers of excellence, but it can also be used to represent organization structures or processes, facilitating communication between the right individuals for specific tasks. Project managers frequently use Gantt charts and waterfall charts to illustrate workflows.

Visual discovery and every day data viz are more closely aligned with data teams. While visual discovery helps data analysts, data scientists, and other data professionals identify patterns and trends within a dataset, every day data viz supports the subsequent storytelling after a new insight has been found. Data visualization is a critical step in the data science process, helping teams and individuals convey data more effectively to colleagues and decision makers. However, it's important to remember that it is a skillset that can and should extend beyond your core analytics team.

7.3. KT0702 - Data visualization using graphics

What is data visualization in graphic design?

Data visualization is a coherent way to visually communicate quantitative content. Depending on its attributes, data may be represented in different ways, such as line graphs and scatter plots.

The objective of your visual

Before making the visualization, it is best to ask yourself what the audience will be looking for in your chart. Understand the requirements and preferences of your viewer. Know their background. Do they have enough time for a detailed visualization? How aware are they of the context of the visualization? What additional information are they looking for? Are they aware of the graphs being used? And so on. Your viewer's information needs should be your guide in creating effective and compelling data visualizations.

Choose the right visualization for your data

There are a tremendous number of charts available. Choosing the right visualization is paramount when you're presenting to a senior leader. It is not easy as it sounds, because an incorrect representation can lead to a wrong message or wrong decision taken by the audience or whatever you've in your mind when you were creating that chart, that message might not be conveyed to the audience. Here, your focus should be on conveying the right message to your audience in an optimal way. Now let me take you through the type of messages, that we usually send out when we're creating impactful visualizations in business.

Column charts

- It is used to compare values across multiple categories.

- Here, the category appears horizontally(X-axis) and values vertically(Y-axis).
- In the column charts, you can also show information about parts of a whole across different categories, and you can show this in absolute value as well as relative terms. Here comes the concept of a stacked column chart and 100% stacked column charts.

Bar charts

- As you're quite familiar with column charts, you will find that working with bar charts is very synonymous.
- The only difference between them is that in a bar chart, values are represented on the X-axis and categories on the Y-axis.
- We typically use a bar graph to show values across categories when the duration or category text is long.
- Stacked bar charts are used to compare parts of a whole(relative and absolute) and compare change over categories or time.

Line charts

- It is one of the most popular charts and vitally used in most industries.
- Whether you're analyzing sales data, whether you're looking at year-on-year profit, whether you're looking at how a person's salary increases in the last year, line charts are very helpful in these scenarios.
- The line chart is used to show trends over time or categories.
- Here, the category appears horizontally(X-axis) and value vertically(Y-axis).

Scatter plots

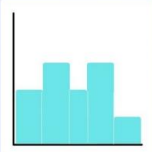
- An XY(Scatter) chart uses numerical values along both axes.
- Scatter plots are useful for showing a correlation between the data points that may not be easy to see from the data alone.
- It is used for displaying and comparing numerical values, such as scientific or statistical data.

Distribution charts

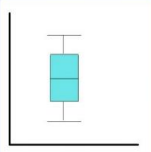
These charts are used to show the spread of the data values over categories or continuous values. We can use the following charts in order to visualize the distribution of the data. For example Distribution of bugs found in 10 weeks of the software testing phase.

Distribution

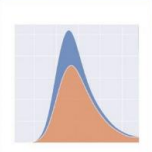
- To show the spread of data values over categorical or continuous values.
- Use the following charts to represent it.



Histogram




Boxplot



KDE plot

- **Example:**
 - Distributions of bugs found in 10 weeks of software testing phase.



- Medians are useful because they're not swayed by outliers as mean is.
- Within the box itself, there is 25% of data above the median and 25% of data below the median, so 50% of the data is within the box.
- By using this plot, we can easily spot outliers and the distribution of the plot.

Histogram

- It is used to graphing the frequency over a distribution. It is a very useful graph in the analytics world and can infer many useful insights from the data.
- Visually, all the bars are touching each other with no space between them.

Box plot

- It is also known as Box and whiskers plot.
- The line in the middle of the box is the median value. This means that 50% of the data are above the median value and 50% of the data are below the median value.

7.4. KT0703 - Data visualization using system for declaratively creating graphics

What is graphic design in simple terms?

Graphic design is the practice of composing and arranging the visual elements of a project. Designing the layout of a magazine, creating a poster for a theatre performance, and designing packaging for a product are all examples of graphic design.

In fact, almost everyone today practices some form of graphic design in their daily life — whether it's adding text to an image for social media or color-coding a spreadsheet for work.

What is the main purpose of graphic design?

The objective of graphic design is to convey or enhance a message.

Good graphic art streamlines communication. Just picture a spreadsheet with data analytics. A graphic designer might use different colors to highlight which metrics are rising and which are dropping, thus making it easier for the viewer to quickly understand what's going well and what needs to be adjusted.

Well executed graphic design can also elicit an emotional response from the viewer or even motivate them to take action. The “sign up” page on a website, for example, is typically designed to entice visitors to join an email list or start a free trial. Meanwhile, food packaging design aims to make the food inside seem more appealing to eat.

7.5. KT0704 - File formats of graphic outputs

What is the file format for graphic?

Definition: Graphic images are stored digitally using a small number of standardized graphic file formats, including bit map, TIFF, JPEG, GIF, PNG; they can also be stored as raw, unprocessed data

Web Graphic Formats

There are three file formats for graphics used on the web: JPG, GIF, and PNG. Each of these file formats are designed with a specific purpose in mind, so it is important to understand the differences when we use them in our websites.

JPG

The JPG image format was designed to efficiently store and compress realistic images and artwork (both in color and greyscale). The JPG format does a very good job of compressing images with lots of colors and gradations in colors. Think of a JPG as a highly compressed photograph.

The JPG format is not capable of saving any transparencies. If transparency is needed in the background of your image, you must choose a different format.

When saving images in the JPG format, you can choose the level of compression to balance the file size and image quality. File size is directly related to the actual size (in pixels) of the image. A larger pixel size will always result in a larger file size.

GIF and PNG

The GIF and PNG image formats use what is called "index-color". They store a minimized color palette in the image file and keys to where those colors should be located in the image. File size for GIF and PNG images is generally related to the number of colors used. Commons numbers of colors are: 2, 4, 8, 16, 32, 64, 128, 256.

The GIF and PNG image formats are ideal for images with flat colors (no gradients) and hard edges. Common examples of these types of images are logos, logotypes, and illustrations without gradients.

Transparency

The GIF and PNG formats also both support transparency. If you need any level of transparency in your image, you must use either a GIF or a PNG.

GIF images (and also PNG) support 1-color transparency. This basically means that you can save your image with a transparent background.

Unit Overview

The following topics are covered in this unit:

- Organizing high throughput data
- Multiple comparison problem
- Family wide error rates
- False discovery rate
- Error rate control procedures
- Bonferroni correction
- q-values
- Statistical modelling
- Hierarchical Models and the basics of Bayesian Statistics

Learning Outcomes

At the end of this unit the student should be able to

- Explore techniques for managing and analyzing high-throughput data sets, crucial in big data scenarios.

8.1. Welcome to the High-throughput Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

8.1.1. Introduction to this Unit

High-throughput data analysis deals with organizing and interpreting large-scale datasets, often generated in fields like genomics, proteomics, and high-dimensional experiments. Proper organization of high-throughput data is critical to ensure accessibility, reproducibility, and effective analysis. This involves preprocessing data, ensuring consistency in formatting, and structuring it into manageable forms, such as matrices or relational databases. Data visualization and exploratory techniques are often employed to identify trends and potential outliers before conducting deeper analyses.

One key challenge in high-throughput studies is the multiple comparison problem, which arises when performing numerous statistical tests simultaneously. Without proper adjustments, the likelihood of false positives increases. To address this, measures like the family-wise error rate (FWER) aim to control the probability of at least one false positive across all tests. Methods such as the Bonferroni correction offer conservative solutions by adjusting significance thresholds based on the number of comparisons, reducing false positives but potentially increasing false negatives.

The false discovery rate (FDR) provides a more balanced approach by controlling the proportion of false positives among significant results. Techniques like q-values and error rate control procedures are used to manage FDR effectively, ensuring a higher sensitivity in high-throughput experiments. Advanced techniques, such as statistical modeling and hierarchical models, further refine the analysis by incorporating data dependencies and varying levels of uncertainty. Bayesian statistics, which provide a probabilistic framework for incorporating prior knowledge and updating beliefs, are particularly useful for hierarchical modeling. By mastering these concepts, researchers can derive meaningful insights while maintaining statistical rigor in high-throughput analyses.

Unit 8 – High-throughput

8.2. KT0801 - Organizing high throughput data

What is high-throughput data?

“High-throughput data”, the information generated in a massive, fast manner by 'omics' technologies - transcriptomics, metabolomics and proteomics - have opened a new era in biomedical research allowing an exponential increase of biomedical discoveries.

What are high-throughput techniques?

While sequencing information has traditionally been elucidated using a low throughput technique called Sanger sequencing, high throughput sequencing (HTS) technologies are capable of sequencing multiple DNA molecules in parallel, enabling hundreds of millions of DNA molecules to be sequenced at a time.

8.3. KT0802 - Multiple comparison problem

What is the Multiple Testing Problem?

If you run a hypothesis test, there's a small chance (usually about 5%) that you'll get a bogus significant result. If you run thousands of tests, then the number of false alarms increases dramatically. For example, let's say you run 10,000 separate hypothesis tests (which is common in fields like genomics). If you use the standard alpha level of 5% (which is the probability of getting a false positive), you're going to get around 500 significant results — most of which will be false alarms. This large number of false alarms produced when you run multiple hypothesis tests is called the multiple testing problem. (Or multiple comparisons problem).

$$P_{(k)} > \frac{\alpha}{m + 1 - k}$$

When you run multiple tests, the p-values have to be adjusted for how many hypothesis tests you are running. In other words, you have to control the Type I error rate (a Type I error is another name for incorrectly rejecting the null hypothesis). There isn't a universally-accepted way to control for the problem of multiple testing.

Why multiple testing is a problem?

What is the Multiple Testing Problem? If you run a hypothesis test, there's a small chance (usually about 5%) that you'll get a bogus significant result. If you run thousands of tests, then the number of false alarms increases dramatically.

What is multiple comparisons fallacy?

(Also Known As: multiple comparisons, multiplicity, multiple testing problem, the look-elsewhere effect) Description: Claiming that unexpected trends that occur through random chance alone in a data set with a large number of variables are meaningful

In statistics, the multiple comparisons, multiplicity or multiple testing problem occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the observed values.

The more inferences are made, the more likely erroneous inferences become. Several statistical techniques have been developed to address that problem, typically by requiring a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made.

8.4. KT0803 - Family wide error rates

What is the Familywise Error Rate?

The familywise error rate (FWE or FWER) is the probability of a coming to at least one false conclusion in a series of hypothesis tests. In other words, it's the probability of making at least one Type I Error. The term "familywise" error rate comes from family of tests, which is the technical definition for a series of tests on data.

What is a family-wise Type I error rate?

In multiple comparison procedures, family-wise type I error is the probability that, even if all samples come from the same population, you will wrongly conclude that at least one pair of populations differ.

In statistics, family-wise error rate (FWER) is the probability of making one or more false discoveries, or type I errors when performing multiple hypotheses tests.

How to Estimate the Family-wise Error Rate

The formula to estimate the family-wise error rate is as follows:

$$\text{Family-wise error rate} = 1 - (1-\alpha)^n$$

where:

α : The significance level for a single hypothesis test

n: The total number of tests

For example, suppose we conduct 5 different comparisons using an alpha level of $\alpha = .05$. The family-wise error rate would be calculated as:

$$\text{Family-wise error rate} = 1 - (1-\alpha)^c = 1 - (1-.05)^5 = 0.2262.$$

In other words, the probability of getting a type I error on at least one of the hypothesis tests is over 22%!

How to Control the Family-wise Error Rate

There are several methods that can be used to control the family-wise error rate, including:

1. The Bonferroni Correction.

Adjust the α value used to assess significance such that:

$$\alpha_{\text{new}} = \alpha_{\text{old}} / n$$

For example, if we conduct 5 different comparisons using an alpha level of $\alpha = .05$, then using the Bonferroni Correction our new alpha level would be:

$$\alpha_{\text{new}} = \alpha_{\text{old}} / n = .05 / 5 = .01.$$

2. The Sidak Correction.

Adjust the α value used to assess significance such that:

$$\alpha_{\text{new}} = 1 - (1-\alpha_{\text{old}})^{1/n}$$

For example, if we conduct 5 different comparisons using an alpha level of $\alpha = .05$, then using the Sidak Correction our new alpha level would be:

$$\alpha_{\text{new}} = 1 - (1-\alpha_{\text{old}})^{1/n} = 1 - (1-.05)^{1/5} = .010206.$$

3. The Bonferroni-Holm Correction.

This procedure works as follows:

Use the Bonferroni Correction to calculate $\alpha_{\text{new}} = \alpha_{\text{old}} / n$.

Perform each hypothesis test and order the p-values from all tests from smallest to largest.

If the first p-value is greater than or equal to α_{new} , stop the procedure. No p-values are significant.

If the first p-value is less than α_{new} , then it is significant. Now compare the second p-value to α_{new} . If it's greater than or equal to α_{new} , stop the procedure. No further p-values are significant.

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit Name 4

Unit Name 5

Next >>

8.5. KT0804 - False discovery rate

What is false discovery rate in statistics?

In technical terms, the false discovery rate is the proportion of all 'discoveries' which are false. When running a classical statistical test, any time a null hypothesis is rejected it can be considered a 'discovery'.

In statistics, the false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. FDR-controlling procedures are designed to control the FDR, which is the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections of the null). Equivalently, the FDR is the expected ratio of the number of false positive classifications (false discoveries) to the total number of positive classifications (rejections of the null). The total number of rejections of the null include both the number of false positives (FP) and true positives (TP). Simply put, $FDR = FP / (FP + TP)$. FDR-controlling procedures provide less stringent control of Type I errors compared to familywise error rate (FWER) controlling procedures (such as the Bonferroni correction), which control the probability of at least one Type I error. Thus, FDR-controlling procedures have greater power, at the cost of increased numbers of Type I errors.

8.6. KT0805 - Error rate control procedures

How do we control for Type I error rate?

One of the most common approaches to minimizing the probability of getting a false positive error is to minimize the significance level of a hypothesis test. Since the significance level is chosen by a researcher, the level can be changed. For example, the significance level can be minimized to 1% (0.01).

How to Avoid a Type I Error?

It is not possible to completely eliminate the probability of a type I error in hypothesis testing. However, there are opportunities to minimize the risks of obtaining results that contain a type I error.

One of the most common approaches to minimizing the probability of getting a false positive error is to minimize the significance level of a hypothesis test. Since the significance level is chosen by a researcher, the level can be changed. For example, the significance level can be minimized to 1% (0.01). This indicates that there is a 1% probability of incorrectly rejecting the null hypothesis.

However, lowering the significance level may lead to a situation wherein the results of the hypothesis test may not capture the true parameter or the true difference of the test.

8.7. KT0806 - Bonferroni correction

The **Bonferroni correction** is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously (since while a given alpha value may be appropriate for each individual comparison, it is not for the set of all comparisons).

In statistics, the Bonferroni correction is a method to counteract the multiple comparisons problem. Bonferroni correction is the simplest method for counteracting this; however, it is a conservative method that gives greater chance of failure to reject a false null hypothesis than other methods, as it ignores potentially valuable information, such as the distribution of p-values across all comparisons (which, if the null hypothesis is correct for all comparisons, is expected to take uniform distribution).

8.8. KT0807 - q-values

What is a significant q-value?

The q value provides a measure of each feature's significance, automatically taking into account the fact that thousands are simultaneously being tested. Suppose that features with q values $\leq 5\%$ are called significant in some genomewide test of significance. This results in a FDR of 5% among the significant features.

8.9. KT0808 - Statistical modelling

A **statistical model** is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population). A statistical model represents, often in considerably idealized form, the data-generating process.

A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables. As such, a statistical model is "a formal representation of a theory"

What is Statistical Modeling and How is it Used?

Statistical modelling is the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.

When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically. Rather than sifting through the raw data, this practice allows them to identify relationships between variables, make predictions about future sets of data, and visualize that data so that non-analysts and stakeholders can consume and leverage it.

"When you analyze data, you are looking for patterns," says Mello. "You are using a sample to make an inference about the whole."

3 Reasons to Learn Statistical Modeling

While data scientists are most often tasked with building models and writing algorithms, analysts also interact with statistical models in their work on occasion. For this reason, analysts who are looking to excel should aim to obtain a solid understanding of what makes these models successful.

"As machine learning and artificial intelligence become more commonplace, more and more companies and organizations are leveraging statistical modeling in order to make predictions about the future based off data," Mello says. "[So] if you work in the area of data analytics, you need to understand how the underlying models work...No matter what kind of analysis you are doing or what kind of data you are working with, you are going to need to use statistical modelling in some way."

Below are some of the benefits that come from having a thorough understanding of statistical modelling.

1. You will be better equipped to choose the right model for your needs.

There are many different types of statistical models, and an effective data analyst needs to have a comprehensive understanding of them all. In each scenario, you should be able to identify not only which model will help best answer the question at hand, but also which model is most appropriate for the data you're working with.

2. You will be better able to prepare your data for analysis.

Data is rarely ready for analysis in its raw form. To ensure your analysis is accurate and viable, the data must first be cleaned up. This cleanup often includes organizing the gathered information and removing "bad or incomplete data" from the sample.

"Before any statistical model can be completed, you need to explore [and], understand the data," says Mello. "If there is no quality [in the data], then you can't really derive any insights from it."

Once you know how various statistical models work and how they leverage data, it will become easier for you to determine what data is most relevant to the question you are trying to answer, as well.

3. You will become a better communicator.

In most organizations, data analysts are required to communicate their findings with two different audiences. The first audience consists of those on the business team who don't need to understand the details of your analysis, but simply want to know the key takeaways. The second audience consists of those who are interested in the more granular details; this group will want both the list of broad conclusions and an explanation of how you reached them.

Having a thorough understanding of statistical modeling can help you better communicate with both of these audiences, as you will be better equipped to reach conclusions and therefore generate better data visualizations, which are helpful in communicating complex ideas to non-analysts. Simultaneously, a complex understanding of how these models work on the backend will allow you to generate and explain those more granular details when necessary.

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit Name 4

Unit Name 5

Next >>

8.10. KT0809 - Hierarchical Models and the basics of Bayesian Statistics

What is a hierarchical model in statistics?

A hierarchical model is a model in which lower levels are sorted under a hierarchy of successively higher-level units. Data is grouped into clusters at one or more levels, and the influence of the clusters on the data points contained in them is taken account in any statistical analysis

Bayesian hierarchical modelling is a statistical model written in multiple levels (hierarchical form) that estimates the parameters of the posterior distribution using the Bayesian method. The sub-models combine to form the hierarchical model, and Bayes' theorem is used to integrate them with the observed data and account for all the uncertainty that is present. The result of this integration is the posterior distribution, also known as the updated probability estimate, as additional evidence on the prior distribution is acquired

Why are Bayesian hierarchical models?

One important benefit of the Bayesian hierarchical approach is that an appropriate choice of priors and model structure allows us to integrate additional model parameters without excessively increasing model complexity.

What is hierarchical Bayesian inference?

The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next. At each step ideally we will know the conditional distributions. The aim is to build a complete model of the data.

Bayesian statistics mostly involves conditional probability, which is the probability of an event A given event B, and it can be calculated using the Bayes rule. The concept of conditional probability is widely used in medical testing, in which false positives and false negatives may occur.

What is Bayesian statistics simple?

“Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the evidence of new data.”

8.11. KT0810 - Exploratory data analysis for high throughput data

What is high throughput data analysis?

Automation and high throughput. High-throughput (HT) analysis is becoming more and more important. It means analysis of dozens, hundreds, or even thousands of samples per day in a given laboratory or on a particular instrument.

What is exploratory data analysis?

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

Why is exploratory data analysis important in data science?

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

High-dimensional data analysis

Unit 9

Unit Overview

The following topics are covered in this unit:

- Mathematical distance
- Dimension reduction
- Singular value decomposition and principal component analysis
- Multiple dimensional scaling plots
- Factor analysis
- Dealing with batch effects
- Clustering
- Heatmaps

Learning Outcomes

At the end of this unit the student should be able to

- Gain insights into the challenges and strategies for analyzing high-dimensional data, including dimensionality reduction techniques.

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit 4 Name

Unit Name 5

Next >>

9.1. Welcome to the High-dimensional data analysis Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

9.1.1. Introduction to this Unit

High-dimensional data analysis addresses the challenges of working with datasets that have a large number of variables compared to observations, such as gene expression profiles or image data. A fundamental concept in this field is mathematical distance, which quantifies the similarity or dissimilarity between data points. Metrics like Euclidean, Manhattan, and cosine distances are commonly used in high-dimensional spaces to identify patterns or groupings.

Dimension reduction techniques play a critical role in simplifying high-dimensional data while retaining meaningful information. Singular Value

Decomposition (SVD) and Principal Component Analysis (PCA) are widely used methods that transform the data into lower-dimensional spaces by capturing the most variance in the dataset. These methods help uncover underlying structures and make visualization more accessible. Multidimensional Scaling (MDS) plots provide another way to represent high-dimensional data in two or three dimensions, preserving distance relationships to facilitate pattern recognition.

Advanced techniques like factor analysis help identify latent variables influencing observed data, while methods for dealing with batch effects mitigate variability caused by external factors, ensuring the validity of results. Clustering algorithms such as k-means and hierarchical clustering group similar data points, revealing inherent structures in the data. Visualization tools like heatmaps enhance the interpretation of clustering results by providing intuitive, color-coded representations of relationships within the data. Together, these techniques form a comprehensive toolkit for extracting insights and making sense of complex high-dimensional datasets.

Unit 9 – High-dimensional data analysis

9.2. KT0901 - Mathematical distance

In mathematics, a metric or distance function is a function that gives a distance between each pair of point elements of a set. A set with a metric is called a metric space. A metric induces a topology on a set, but not all topologies can be generated by a metric. A topological space whose topology can be described by a metric is called metrizable.

One important source of metrics in differential geometry are metric tensors, bilinear forms that may be defined from the tangent vectors of a differentiable manifold onto a scalar. A metric tensor allows distances along curves to be determined through integration, and thus determines a metric.

Distance is a numerical measurement of how far apart objects or points are. In Physics or everyday usage, distance may refer to a physical length or an estimation based on other criteria (e.g. "two counties over"). The distance from a point A to a point B is sometimes denoted as $|AB|$. In most cases, "distance from A to B" is interchangeable with "distance from B to A". In mathematics, a distance function or metric is a generalization of the concept of physical distance; it is a way of describing what it means for elements of some space to be "close to", or "far away from" each other. In psychology and social sciences, distance is a non-numerical measurement; Psychological distance is defined as "the different ways in which an object might be removed from" the self along dimensions such as "time, space, social distance, and hypotheticality.

9.3. KT0902 - Dimension reduction

What do you mean by dimension reduction?

Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset. More input features often make a predictive modelling task more challenging to model, more generally referred to as the curse of dimensionality.

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension. Working in high-dimensional spaces can be undesirable for many reasons; raw data are often sparse as a consequence of the curse of dimensionality, and analyzing the data is usually computationally intractable (hard to control or deal with). Dimensionality reduction is common in fields that deal with large numbers of observations and/or large numbers of variables, such as signal processing, speech recognition, neuroinformatics, and bioinformatics.

Methods are commonly divided into linear and nonlinear approaches. Approaches can also be divided into feature selection and feature extraction. Dimensionality reduction can be used for noise reduction, data visualization, cluster analysis, or as an intermediate step to facilitate other analyses.

9.4. KT0903 - Singular value decomposition and principal component analysis

What is the difference between SVD and PCA?

The main difference between The Singular value decomposition and principal component analysis is that The SVD is a data-driven Fourier transform generalization, whereas PCA allows us to represent statistical variations in our data sets using a hierarchical coordinate system based on data.

What is SVD in principal component analysis?

Singular Value Decomposition is a matrix factorization method utilized in many numerical applications of linear algebra such as PCA. This technique enhances our understanding of what principal components are and provides a robust computational framework that lets us compute them accurately for more datasets.

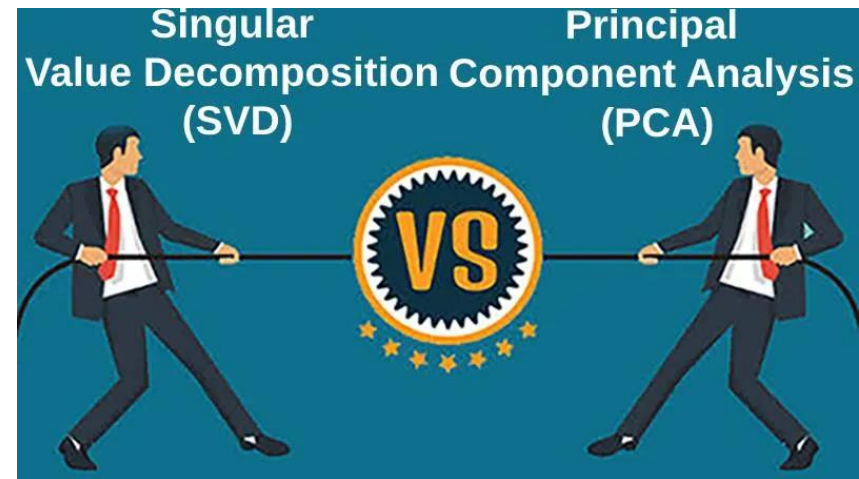
Difference Between Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) (With Table)

The singular value decomposition (SVD) is among the most extensively used and all-purpose helpful features in numerical linear algebra for data acquisition, whereas principal component analysis (PCA) is a well-established method that has introduced a lot of theories about statistics. In particular, PCA provides us with a data-driven hierarchical coordinate system.

Singular Value Decomposition (SVD) vs Principal Component Analysis (PCA)

The main difference between The Singular value decomposition and principal component analysis is that The SVD is a data-driven Fourier transform generalization, whereas PCA allows us to represent statistical

variations in our data sets using a hierarchical coordinate system based on data.



The singular value decomposition (SVD) is the most extensively used feature in numerical linear algebra. It aids in the reduction of data into the key features required for analysis, understanding, and description. The svd is one of the first elements in most data preprocessing and machine learning algorithms for data reduction in particular. The SVD is a data-driven Fourier transform generalization.

The principal component analysis (PCA) is now a statistical tool that has spawned numerous ideas. This will allow us to use a hierarchical set of points to express statistical changes. PCA is a statistical/machine intelligence technique used to determine the major data patterns that maximize overall variation. So the maximum variance is captured by a coordinate system depending on the data's directions.

Comparison Table Between Singular Value Decomposition (SVD) and Principal Component Analysis (PCA)

Parameters of Comparison	Singular Value Decomposition (SVD)	Principal Component Analysis (PCA)
Requirements	Abstract mathematics, matrix decomposition, and quantum physics all require SVD.	Statistics are particularly effective in PCA for analyzing data from the research.
Expression	Factoring algebraic expressions.	similar to approximating factorized expressions.
Methods	It is a method in abstract mathematics and matrix decomposition.	It is a method in Statistics/Machine Learning.
Branch	Helpful in the branch of mathematics.	Helpful in the branch of mathematics.
Invention	The SVD was invented by Eugenio Beltrami and Camille Jordan.	The PCA was invented by Karl Pearson.

What is Singular Value Decomposition (SVD)?

The SVD is strongly linked to the part of a positive definite Matrix's eigenvalue and eigenvector factorization. Although not all matrices may be factorized as PT , any $m \times n$ matrix A can be factorized by permitting it on the left and PT on the right to be any two orthogonal matrices U and V (not necessarily transpose of each other) This type of special factorization is known as SVD.

The sine and cosine expansions are used in all mathematics to approximate functions, and FT is one of the most useful transformations. There are also Bessel and Airy functions, as well as spherical harmonics. And, in the previous generation of computer science and engineering, this mathematical model mathematical transformation was used to transfer a system of interest into a new coordinate system.

What is Principal Component Analysis (PCA)?

PCA is a well-established method that has introduced a lot of theories about statistics. It is equivalent to approximating a factorized statement by maintaining the 'largest' terms and eliminating all smaller' terms. It is a well-

established method that has introduced a lot of theories about statistics. In particular, PCA provides us with a data-driven hierarchical coordinate system.

Principal component analysis (PCA) is often referred to as appropriate orthogonal decomposition. PCA is a method for identifying patterns in data by defining them in terms of similarities and differences. In PCA, there is a data matrix X that contains a collection of measurements from different experiments, and two independent experiments are represented as large row factors at x_1, x_2 , and so on.

9.5. KT0904 - Multiple dimensional scaling plots

Multidimensional scaling (MDS) is a technique that creates a map displaying the relative positions of a number of objects, given only a table of the distances between them. The map may consist of one, two, three, or even more dimensions. The program calculates either the metric or the non-metric solution

What is multidimensional scaling with example?

For example, given a matrix of perceived similarities between various brands of air fresheners, MDS plots the brands on a map such that those brands that are perceived to be very similar to each other are placed near each other on the map, and those brands that are perceived to be very different from each other are

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. MDS is used to translate "information about the pairwise 'distances' among a set of objects or individuals" into a configuration of $\{ \}$ points mapped into an abstract Cartesian space.

More technically, MDS refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. It is a form of non-linear dimensionality reduction.

Given a distance matrix with the distances between each pair of objects in a set, and a chosen number of dimensions, N , an MDS algorithm places each object into N -dimensional space (a lower-dimensional representation) such that the between-object distances are preserved as well as possible. For $N = 1, 2,$ and 3 , the resulting points can be visualized on a scatter plot.

9.6. KT0905 - Factor analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors.

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved (underlying) variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors plus "error" terms, hence factor analysis can be thought of as a special case of errors-in-variables models.

Simply put, the factor loading of a variable quantifies the extent to which the variable is related to a given factor.

A common rationale behind factor analytic methods is that the information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Factor analysis is commonly used in psychometrics, personality psychology, biology, marketing, product management, operations research, finance, and machine learning. It may help to deal with data sets where there are large numbers of observed variables that are thought to reflect a smaller number of underlying/latent variables. It is one of the most commonly used inter-dependency techniques and is used when the relevant set of variables shows a systematic inter-dependence and the objective is to find out the latent factors that create a commonality.

9.7. KT0906 - Dealing with batch effects

In molecular biology, a batch effect occurs when non-biological factors in an experiment cause changes in the data produced by the experiment. Such effects can lead to inaccurate conclusions when their causes are correlated with one or more outcomes of interest in an experiment. They are common in many types of high-throughput sequencing experiments, including those using microarrays, mass spectrometers,[1] and single-cell RNA-sequencing data. They are most commonly discussed in the context of genomics and high-throughput sequencing research, but they exist in other fields of science as well.

9.8. KT0907 - Clustering

Why clustering is used?

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

Types of Clustering

Broadly speaking, clustering can be divided into two subgroups :

- **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.
- **Soft Clustering:** In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each customer is assigned a probability to be in either of 10 clusters of the retail store.

Types of clustering algorithms

Since the task of clustering is subjective, the means that can be used for achieving this goal are plenty. Every methodology follows a different set of rules for defining the 'similarity' among data points. In fact, there are more than 100 clustering algorithms known. But few of the algorithms are used popularly, let's look at them in detail:

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

9.9. KT0908 - Heatmaps

What heatmap means?

A heatmap is a graphical representation of data that uses a system of color-coding to represent different values. Heatmaps are used in various forms of analytics but are most commonly used to show user behavior on specific webpages or webpage templates.

What are heatmaps used for?

Heatmaps are used to show relationships between two variables, one plotted on each axis. By observing how cell colors change across each axis, you can observe if there are any patterns in value for one or both variables

Basic machine learning and artificial intelligence concepts

Unit 10

Unit Overview

The following topics are covered in this unit:

- ML concepts and principles
- ML application
- ML technologies
- Supervised learning
- Unsupervised learning
- Reinforcement learning
- Algorithms

Learning Outcomes

At the end of this unit the student should be able to

- Get introduced to the basics of machine learning and AI, understanding their application in data analysis and prediction.

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit 4 Name

Unit Name 5

Next >>

10.1. Welcome to the Basic machine learning and artificial intelligence concepts Unit

Welcome in this unit, you will explore key concepts and practical applications.

To make the most of your learning experience, follow these instructions:

- Review: Start by reading through the provided materials for this unit. Pay attention to the key ideas and concepts presented.
- Study your prescribed material
- Follow the Study Material References: Utilise the references to the prescribed book(s) to delve deeper into the subject matter. These study materials will enhance your understanding and provide insights.

Feel free to engage in discussions, ask questions, and collaborate with your peers on the StudentHub to deepen your understanding of this unit.

Enjoy your learning journey!

10.1.1. Introduction to this Unit

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that focuses on building systems capable of learning and improving from experience without explicit programming. ML concepts and principles revolve around using algorithms to identify patterns in data, make predictions, and improve performance over time. Core principles include data preparation, feature engineering, model selection, and evaluation to ensure accurate and efficient learning.

ML applications span diverse fields, such as healthcare (predicting disease outcomes), finance (fraud detection), and retail (recommendation systems). These applications rely on ML technologies, including software frameworks like TensorFlow, PyTorch, and Scikit-learn, which provide robust tools for building and deploying ML models. The choice of technology often depends on the complexity of the problem and the scalability requirements.

ML is categorized into three main types: supervised learning, where models are trained on labeled data to predict outputs (e.g., classification and regression tasks); unsupervised learning, which identifies patterns or structures in unlabeled data, such as clustering or dimensionality reduction; and reinforcement learning, where agents learn optimal actions by interacting with an environment and receiving feedback in the form of rewards or penalties. Each type employs different algorithms, such as decision trees, support vector machines, k-means clustering, or deep learning models, tailored to solve specific problems. By mastering these foundational concepts, practitioners can leverage ML to develop intelligent systems that adapt and evolve with data.

Unit 10 – Basic machine learning and artificial intelligence concepts

10.2. KT1001 - ML concepts and principles

What are ml concepts?

Machine learning is the way to make programming scalable. Traditional Programming: Data and program is run on the computer to produce the output. Machine Learning: Data and output is run on the computer to create a program. This program can be used in traditional programming.

Traditional Programming



Machine Learning



10.3. KT1002 - ML application

What is ML and its application?

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values

What application is used for machine learning?

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion: Facebook provides us a feature of auto friend tagging suggestion.

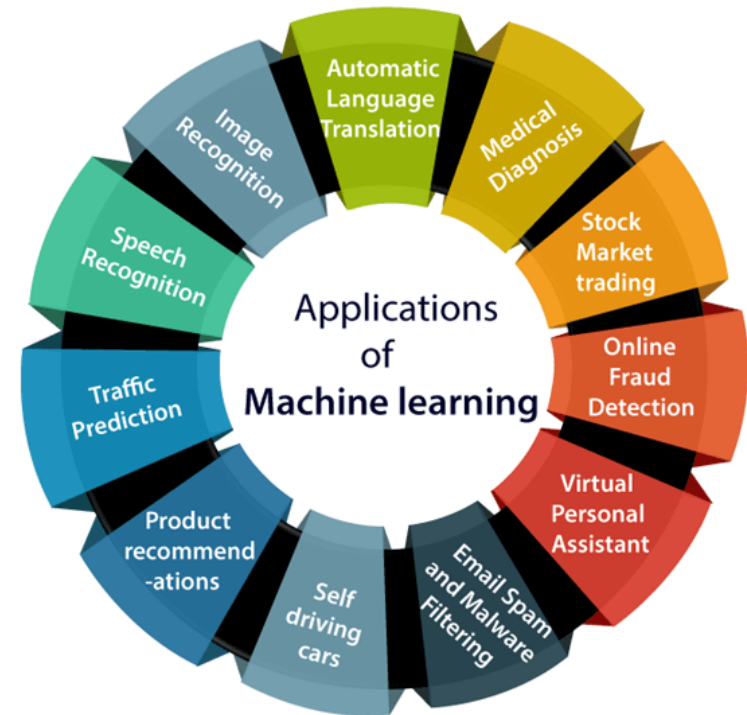
How does supervised machine learning work?

Supervised machine learning requires the data scientist to train the algorithm with both labelled inputs and desired outputs. Supervised learning algorithms are good for the following tasks:

- **Binary classification:** Dividing data into two categories.
- **Multi-class classification:** Choosing between more than two types of answers.
- **Regression modeling:** Predicting continuous values.
- **Ensembling:** Combining the predictions of multiple machine learning models to produce an accurate prediction.

Applications of Machine learning

Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day. We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc. Below are some most trending real-world applications of Machine Learning:



1. Image Recognition:

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion:

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

It is based on the Facebook project named "Deep Face," which is responsible for face recognition and person identification in the picture.

2. Speech Recognition

While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition." At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

3. Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

Real Time location of the vehicle from Google Map app and sensors

Average time has taken on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

4. Product recommendations:

Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

5. Self-driving cars:

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

6. Email Spam and Malware Filtering:

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail:

- Content Filter
- Header filter
- General blacklists filter
- Rules-based filters
- Permission filters

Some machine learning algorithms such as Multi-Layer Perceptron, Decision tree, and Naïve Bayes classifier are used for email spam filtering and malware detection.

7. Virtual Personal Assistant:

We have various virtual personal assistants such as Google assistant, Alexa, Cortana, Siri. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.

These virtual assistants use machine learning algorithms as an important part.

These assistant record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

8. Online Fraud Detection:

Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts, fake ids, and steal money in the middle of a transaction. So to detect this, Feed Forward Neural network helps us by checking whether it is a genuine transaction or a fraud transaction.

For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

9. Stock Market trading:

Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine

learning's long short term memory neural network is used for the prediction of stock market trends.

10. Medical Diagnosis:

In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain.

It helps in finding brain tumors and other brain-related diseases easily.

11. Automatic Language Translation:

Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.

10.4. KT1003 - ML technologies

Who's using machine learning and what's it used for?

Today, machine learning is used in a wide range of applications. Perhaps one of the most well-known examples of machine learning in action is the recommendation engine that powers Facebook's news feed.

Facebook uses machine learning to personalize how each member's feed is delivered. If a member frequently stops to read a particular group's posts, the recommendation engine will start to show more of that group's activity earlier in the feed.

Behind the scenes, the engine is attempting to reinforce known patterns in the member's online behavior. Should the member change patterns and fail to read posts from that group in the coming weeks, the news feed will adjust accordingly.

In addition to recommendation engines, other uses for machine learning include the following:

- Customer relationship management. CRM software can use machine learning models to analyze email and prompt sales team members to respond to the most important messages first. More advanced systems can even recommend potentially effective responses.
- Business intelligence. BI and analytics vendors use machine learning in their software to identify potentially important data points, patterns of data points and anomalies.
- Human resource information systems. HRIS systems can use machine learning models to filter through applications and identify the best candidates for an open position.

- Self-driving cars. Machine learning algorithms can even make it possible for a semi-autonomous car to recognize a partially visible object and alert the driver.
- Virtual assistants. Smart assistants typically combine supervised and unsupervised machine learning models to interpret natural speech and supply context.

What are the advantages and disadvantages of machine learning?

Machine learning has seen use cases ranging from predicting customer behavior to forming the operating system for self-driving cars.

When it comes to advantages, machine learning can help enterprises understand their customers at a deeper level. By collecting customer data and correlating it with behaviors over time, machine learning algorithms can learn associations and help teams tailor product development and marketing initiatives to customer demand.

Some companies use machine learning as a primary driver in their business models. Uber, for example, uses algorithms to match drivers with riders. Google uses machine learning to surface the ride advertisements in searches.

But machine learning comes with disadvantages. First and foremost, it can be expensive. Machine learning projects are typically driven by data scientists, who command high salaries. These projects also require software infrastructure that can be expensive.

There is also the problem of machine learning bias. Algorithms trained on data sets that exclude certain populations or contain errors can lead to inaccurate models of the world that, at best, fail and, at worst, are discriminatory. When an enterprise bases core business processes on biased models it can run into regulatory and reputational harm.

How to choose the right machine learning model

The process of choosing the right machine learning model to solve a problem can be time consuming if not approached strategically.

- **Step 1:** Align the problem with potential data inputs that should be considered for the solution. This step requires help from data scientists and experts who have a deep understanding of the problem.
- **Step 2:** Collect data, format it and label the data if necessary. This step is typically led by data scientists, with help from data wranglers.
- **Step 3:** Chose which algorithm(s) to use and test to see how well they perform. This step is usually carried out by data scientists.
- **Step 4:** Continue to fine tune outputs until they reach an acceptable level of accuracy. This step is usually carried out by data scientists with feedback from experts who have a deep understanding of the problem.

Importance of human interpretable machine learning

Explaining how a specific ML model works can be challenging when the model is complex. There are some vertical industries where data scientists have to use simple machine learning models because it's important for the business to explain how every decision was made. This is especially true in industries with heavy compliance burdens such as banking and insurance.

10.5. KT1004 - Supervised learning

How does supervised machine learning work?

Supervised machine learning requires the data scientist to train the algorithm with both labeled inputs and desired outputs. Supervised learning algorithms are good for the following tasks:

- **Binary classification:** Dividing data into two categories.
- **Multi-class classification:** Choosing between more than two types of answers.
- **Regression modeling:** Predicting continuous values.
- **Ensembling:** Combining the predictions of multiple machine learning models to produce an accurate prediction.

How does semi-supervised learning work?

Semi-supervised learning works by data scientists feeding a small amount of labelled training data to an algorithm. From this, the algorithm learns the dimensions of the data set, which it can then apply to new, unlabelled data. The performance of algorithms typically improves when they train on labelled data sets. But labelling data can be time consuming and expensive. Semi-supervised learning strikes a middle ground between the performance of supervised learning and the efficiency of unsupervised learning. Some areas where semi-supervised learning is used include:

Machine translation: Teaching algorithms to translate language based on less than a full dictionary of words.

Fraud detection: Identifying cases of fraud when you only have a few positive examples.

10.6. KT1005 - Unsupervised learning

How does unsupervised machine learning work?

Unsupervised machine learning algorithms do not require data to be labeled. They sift through unlabeled data to look for patterns that can be used to group data points into subsets. Most types of deep learning, including neural networks, are unsupervised algorithms. Unsupervised learning algorithms are good for the following tasks:

- **Clustering:** Splitting the dataset into groups based on similarity.
- **Anomaly detection:** Identifying unusual data points in a data set.
- **Association mining:** Identifying sets of items in a data set that frequently occur together.
- **Dimensionality reduction:** Reducing the number of variables in a data set.

10.7. KT1006 - Reinforcement learning

How does reinforcement learning work?

Reinforcement learning works by programming an algorithm with a distinct goal and a prescribed set of rules for accomplishing that goal. Data scientists also program the algorithm to seek positive rewards -- which it receives when it performs an action that is beneficial toward the ultimate goal -- and avoid punishments -- which it receives when it performs an action that gets it farther away from its ultimate goal. Reinforcement learning is often used in areas such as:

- **Robotics:** Robots can learn to perform tasks the physical world using this technique.
- **Video gameplay:** Reinforcement learning has been used to teach bots to play a number of video games.
- **Resource management:** Given finite resources and a defined goal, reinforcement learning can help enterprises plan out how to allocate resources.

10.8. KT1007 - Algorithms

Are ML models algorithms?

Machine learning models are output by algorithms and are comprised of model data and a prediction algorithm. Machine learning algorithms provide a type of automatic programming where machine learning models represent the program.

How do you write an ML algorithm?

6 Steps To Write Any Machine Learning Algorithm From Scratch: Perceptron Case Study

1. Get a basic understanding of the algorithm.
2. Find some different learning sources.
3. Break the algorithm into chunks.
4. Start with a simple example.
5. Validate with a trusted implementation.
6. Write up your process.

Unit Overview

In this unit, the following resources will be included:

- The sources referenced in this content is listed below

Content Sources

- <https://www.simplilearn.com> > ... > AI & Machine Learning
- In <https://builtin.com> > data-science > tour-top-

Welcome

Unit Name 1

Unit Name 2

Unit Name 3

Unit Name 4

Unit Name 5

Next >>