Programs Offered

Post Graduate Programmes (PG)

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

Diploma Programmes

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

Undergraduate Programmes (UG)

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

AMITY UNIVERSITY | DIRECTORATE OF DISTANCE & ONLINE EDUCATION
Amity Helpline: 1800-102-3434 (Toll-free), 0120-4614200
For Distance Learning Programmes: dladmissions@amity.edu | www.amity.edu/addoe
For Online Learning programmes: elearning@amity.edu | www.amityonline.com

Product code
81030112001159

AMITY

Introduction to Data Science

Introduction to Data Science

AMITY UNIVERSITY | DIRECTORATE OF DISTANCE & ONLINE EDUCATION

Introduction to Data Science

AMITY UNIVERSITY | DIRECTORATE OF DISTANCE & ONLINE EDUCATION

Published by Amity University Press for exclusive use of Amity Directorate of Distance and Online Education, Amity University, Noida-201313

Contents

Module - I: Understanding Data Science

Learning Objectives

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

- Understand the basic of data science
- Discuss how data science is different from computer science
- Describe use of data science
- Know different stages of data science
- Infer struggle and challenges of data science
- Explain process of data science
- Summarise concepts to big data
- Recognise utilisation of machine learning
- Discuss opportunity in data science

Introduction

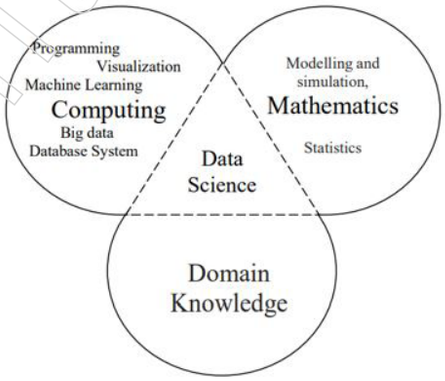
The growth of the Internet and communication technology over the past ten years has produced a significant volume of unstructured data. The usage of social media and mobile technology by people has led to the generation of unstructured data, which includes unformatted textual, image, audio and video data, among other types of data.

Data Science Basics

Data science is a branch of study that combines subject-matter knowledge, computer proficiency and knowledge of mathematics and statistics to draw valuable conclusions from data. By using machine learning techniques to a variety of data types, such as numbers, text, photos, videos and audio, data scientists create artificial intelligence (AI) systems that can do jobs that typically need human knowledge.

1. What is Data Science?

Data science is a multidisciplinary science that aims to analyse data to provide information that can be applied to making decisions. This information may take the shape of forecasting models, predictive planning models, or other models that use comparable patterns.



Application of data science

The following are some of the areas in which data science can be useful.

Notes

- ❖ Making business decisions like assessing the viability of organisations they intend to work with is made easier with its assistance.
- ❖ It may help in developing more accurate projections for the future, such as helping a corporation develop strategic plans based on current patterns.
- ❖ It may find patterns in different data sets that are similar, leading to applications like fraud detection, targeted advertising, etc. Data science is generally a step in the right direction for commercial decision-making, especially in the modern world where data is being produced at a rate of Zetta bytes.

2. Reasons for Data Science

When building models and making predictions, data scientists commonly employ algorithms and other techniques. Data scientists operate in the subject of data science. They often make use of artificial intelligence, especially its subfields of machine learning and deep learning.

Did you know that Southwest Airlines was once able to use data to save $100 million? They could adjust how their resources were used by reducing the amount of time their jets sat idle on the tarmac. In conclusion, no business today could possibly envision a world without data.

3. History of Data Science

Early in the 1960s, the phrase “Data Science” was coined to designate a new profession that would aid in the comprehension and analysis of the massive volumes of data that were being gathered at the time. (The genuinely vast volumes of data that would be produced over the ensuing 50 years were impossible to predict at the time.)

Data science is a constantly developing field. To gather data and create intelligent forecasts across a variety of sectors, it employs computer science and statistical approaches. Data science is utilised in business to assist in decision-making, as well as other industries like astronomy and medicine.

A functional data scientist is skilled in many programming languages and has a thorough understanding of software architecture, unlike a traditional statistician. The data scientist defines the issue, selects the key data sources and develops a framework for collecting and screening the necessary data. Software often handles data collecting, processing and modelling. To get a deeper understanding of the data assets being examined, they employ the data science principles and all relevant subfields and practises.

One odd—and potentially harmful—outcome of the Data Science revolution has been a trend towards developing programming that is steadily becoming more conservative. Data scientists have been found to spend too much time and effort on too complex algorithms when simpler ones would suffice. As a result, abrupt “innovative” changes occur far less frequently.

As a result, many data scientists now strive to divide concepts into smaller portions because they believe that making sweeping adjustments is simply too dangerous. Each component is examined before being gradually phased into the data flow. While more conservative programming is quicker and more effective, experimentation is minimised and fresh, “outside-the-box” ideas and discoveries are restricted.

4. Fundamentals of Data Science

Complex topics like data science, data analytics and business analytics involve

Notes

a wide range of ideas from arithmetic, statistics, programming, computing and management level abilities. The five core concepts of data science are listed below.

1. Data Science Concept: Machine Learning

   A system is programmed to complete a specific task automatically in the branch of artificial intelligence known as machine learning. Afterward, the system uses data to self-learn, spot patterns and make judgements with little to no human input.

2. Data Science Concept: Algorithms

   Algorithms are a particular collection of guidelines or procedures that are applied in a computation to address issues or complete a task. The simplest algorithm, for instance, is a recipe - a list of instructions to follow to get a particular result.

3. Data Science Concept: Statistical Models

   Mathematical models known as statistical models describe the connections between random and non-random variables. By mathematically modelling observable data, datasets are analysed to draw conclusions from the presented samples. This field's central idea is the data science concept. Based on the data that is currently available, models can be used to extract information or anticipate likely outcomes.

4. Data Science Concept: Regression Analysis

   Statistical analysis is done using regression that evaluates relationships between a dependent variable and independent variables to provide a real numerical value that reflects a quantity on a line. Examples include temperature and sales turnover.

5. Data Science Concept: Programming

Python, R for statistics and SQL for database construction and management are all frequently used programming languages in data science. However, people who specialise solely in data interpretation and analysis or in business analytics do not frequently study these languages.

1. Data Science Mindset

   Data Science is a discipline that integrates math, programming and visualisation approaches and applies scientific methods to commercial domains or challenges, such as forecasting future customer behaviour, scheduling air traffic routes, or identifying voice patterns.

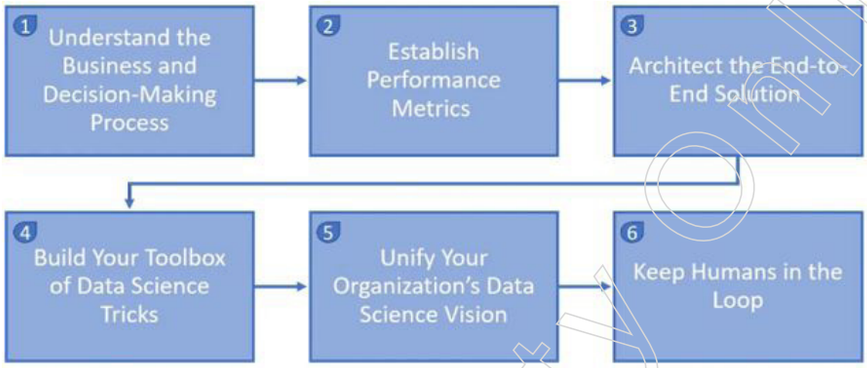
   To be a data-driven organisation, a team of data scientists must be fully integrated into the business and the operational foundation of the organisation (techniques, procedures, infrastructures and culture) must be modified.

   The Healthy Data Science Organisation Framework is a collection of approaches, tools and resources that, when properly used, may help your organisation become more data driven across the board, including in the areas of business understanding, data creation and acquisition, modelling, model deployment and management. This framework includes six key principles:

   1. Understand the Business and Decision-Making Process
   2. Establish Performance Metrics

Notes

3. Architect the End-to-End Solution
4. Build Your Toolbox of Data Science Tricks
5. Unify Your Organisation's Data Science Vision
6. Keep Humans in the Loop



Principle of Data Science

By incorporating these six recommendations from the Organisation Framework for Healthy Data Science into the data analysis process, organisations can improve their business decisions. Their conclusions will be backed up by information that has been carefully obtained and analysed.

Difference Between "Computer Science" and "Data Science"

The study of computers and computer-related ideas is known as computer science. It is essentially the study of how processes and data from programmes interact. It deals with the application of different algorithms to information manipulation. Therefore, the study of hardware, software and other components like networking and the internet is the main focus of computer science.

Data Science Vs Computer Science

Computer science is the study of computers and computer-related ideas. It is essentially the study of how processes and data from programmes interact. It deals with the application of different algorithms to information manipulation. Computer science therefore concentrates on the investigation of both hardware and software also other elements, such as networking and the internet.

The study of computer architecture and operation is covered in the hardware area of computer science. The field of computer science called software includes the study of programming concepts and languages. Operating systems and compilers are other topics covered in computer science.

The study of computers and computing systems is known as computer science,

Notes

including its design, implementation and applications. It covers a broad range of topics, including computer architecture, databases, operating systems, programming languages, algorithms, data structures and software engineering.

The study of how to extract knowledge and information from data using various scientific approaches, algorithms and procedures is known as data science. In this respect, it can be viewed as a collection of many mathematical instruments, algorithms, statistics and machine learning methods that are used to unearth obfuscated patterns and insights from data to support decision-making.

Both organised and unstructured data are dealt with in data science. Both data mining and big data are relevant to it. Data science entails analysing historical trends, using the findings to reframe current trends and forecast future trends.

The interdisciplinary subject of data science, on the other hand, combines statistics, mathematics and computer science to draw conclusions and information from data. It entails gathering, handling, interpreting and visualising facts to comprehend complex events and arrive at wise conclusions. To get insights from data, data scientists employ a variety of tools and approaches, including machine learning, data mining, statistical modelling and data visualisation.

Data science is a specialised area that employs computing technology to draw conclusions and information from data, whereas computer science is a broad field that covers the study of computing technology and its applications.

Since data science and computer science have diverse areas of emphasis, each of these fields of technology offers a variety of job opportunities. Data gathering and analytics specialisations are part of data science jobs. These positions emphasise working with and understanding data from a company to assist employers in making decisions. Some data science roles include:

- Data scientist
- Database administrator
- Data Architect
- Data analyst
- Statistician

The roles in computer science vary depending on the skills and interests of a candidate, but most positions in this field enhance technology and create or maintain digital products for a business. Some occupations in computer science also include the physical components and hardware of a computer system. Computer science jobs include:

- Software engineer
- Web developer
- IT analyst
- Computer programmer
- UI and UX specialist

Applications of Data Science

Given below are the main applications of data science:

- Smart Recommendations: Applications for data science examine vast amounts of historical data and offer useful recommendations.

Notes

- ❖ Big Data Analytics: Applications for data science analyse vast amounts of data, including sensor data, IoT data, social media data, log data and more, to draw conclusions.
- ❖ Web Search: Analyse the data from the web to produce precise search results. Data Analytics: By analysing and visualising data, you can gather knowledge that will enable us to generate data-driven forecasts, predictions and recommendations.
- ❖ Business Intelligence: Business intelligence is the primary usage of data science applications. Organisations can use data science-based business analytics applications to identify challenges, spot cross-sell and up-sell opportunities, assess sales offers, investigate marketing opportunities, product pricing, cost-saving strategies, perform ROI (return on investment) analysis, forecast revenue loss, plan stores, analyse customer data and perform seasonality analysis, among other things.
- ❖ Healthcare: For active monitoring and to lessen health issues, smart data is used.
- ❖ Finance: Numerous financial applications, including fraud detection, risk identification, data-driven insights, pattern recognition, predictive analytics and others, can benefit from the use of data science techniques.
- ❖ Automobile: Using the training data to programme the driverless cars and assessing the models' performance.
- ❖ Supply Chain and Inventory Management: Use data science to estimate demand, revenue and other outcomes.
- ❖ Telecom: Utilise data science techniques to comprehend forecasts for customer turnover. Election campaigning, chatbots, virtual assistants, automated vehicles, disaster prediction, product pricing, preventive maintenance, customer retention, pattern recognition, online advertisements, demand forecasting, trend analysis, determining the effectiveness of campaigns, recommendations and other uses for data science are just a few examples.

Through exploratory data analysis, you can describe the events, phenomenon and scenarios. To diagnose the cause of the incident, you can also examine the data dimensions. Following data collection, analysis and model construction, it is possible to forecast future events and suggest actions that can be taken to either hasten or avert them. Given below are the popular applications of data science:

- ❖ Can also look at the data dimensions to determine the incident's root cause. It is possible to predict future occurrences and identify actions that might be taken to either hasten or avert them after data collecting, analysis and model creation.
- ❖ Identifying the most pertinent product recommendations based on a user's past purchases, browsing patterns and interests of other users.
- ❖ Identifying an email as spam using attributes that are constantly being learned.
- ❖ Real-time detection of barriers by autonomous vehicles.
- ❖ Recognise customer behaviours immediately.
- ❖ The likelihood of client attrition can be predicted using the signs.

1. Many Paths to Data Science: A degree in computer science is not usually the first step on the road to data science. Many aspiring data scientists begin their careers in math-intensive disciplines like physics, chemistry, engineering, or statistics. Even in fields

Notes

where a degree in data science is currently offered separately, it is frequently housed in a department of mathematics or engineering. Others are physical science-focused multidisciplinary programmes or are extracurricular choices for (often) mathematics majors.

2. Advice for Data Scientists: Data scientists must exercise creativity to come up with fresh solutions to issues. This calls for creative problem-solving and thinking beyond the box. Additionally, it's crucial to be able to articulate your thoughts clearly so that others may comprehend them.

    1. Statistical thinking: Statistical knowledge is a vital component of our toolset since data scientists are experts at turning data into information. Knowing your algorithms and when to utilise them is perhaps the most crucial part of a data scientist's work. However, mastering it can require both art and science.
    2. Technical acumen: Data scientists work in teams and write code to develop tools, pipelines, packages, modules, features, dashboards and other things. Both the front end and the back end have their own code. Can perform structured and unstructured work. When can't get the answer as needed, can "roll our own" tools by sifting through new formats and legacy code.
    3. Multi-modal communication skills: A clever data scientist can contextualise and convey a problem and its solution to interested people from a variety of backgrounds using metaphor, common ground, acute listening and storytelling. This includes vocal communication for project requirements, check-in meetings, iterative design and presentations, as well as written communication for statements of work or reports. Visualisation and story clarity through graphic communication.
    4. Curiosity: A data scientist is an expert in statistics and technology who can explain the Markov chain to a checker at the grocery store. What else makes the elite unique? Curiosity is the first of our three crucial soft talents. Many people who are drawn to data science find the chance to work on a never-ending stream of innovative and difficult issues to be particularly enticing.
    5. Creativity: The best data scientists, however, go beyond aesthetics and communication to solve problems creatively and have an odd relationship with the word "no." The word "no" irritates great data scientists to the point where they find a means to get around it, over it, or through it, or they back off and choose another course of action.
    6. Grit: A reasonable individual could take an unforeseen leave of absence due to any difficulties or obstacles. Grit is that inner motivation that pushes us beyond hurdles, transforms setbacks into design restrictions, keeps us moving forward in the face of true failure, aids us in restraining the need to take things personally and wipes the dirt off our shoulders. When grit is in action, you will be less competitive, allowing us to support and educate one another.

3. A Day in the Life of a Data Scientist: Data scientists need both human understanding and statistical and machine learning techniques and approaches to be able to extract and interpret relevant information from data. She spends a lot of time gathering, cleaning and munging data because no data is ever completely clean.

As might be assumed given the nature of the job, data is essential to a data scientist's daily tasks. Data scientists invest a lot of time in gathering, analysing and

Notes

modifying data, but they do so in a variety of methods and with a variety of objectives. Among the data-related duties a data scientist might take on are:

- ❖ Pulling data
- ❖ Merging data
- ❖ Analysing data
- ❖ Looking for patterns or trends
- ❖ Using a wide variety of tools, including R, Tableau, Python, MATLAB, Hive, Impala, PySpark, Excel, Hadoop, SQL and/or SAS
- ❖ Developing and testing new algorithms
- ❖ Trying to simplify data problems
- ❖ Developing predictive models
- ❖ Building data visualisations
- ❖ Writing up results to share with others
- ❖ Pulling together proofs of concepts

- Stages and Data Science Elements

Workflows for data science commonly occur in a variety of fields and specialties, including biology, geography, finance and business, among others. Consequently, very different approaches and data sets can be employed in Data Science initiatives with very varied goals and challenges.

- International Risks of Data Science- Old and New

Any company project entails a variety of risks, if they are ignored, the team may not be able to produce the desired results. Every data science project has its own unique challenges, therefore it's important to be ready for them.

The value of knowledge has surpassed that of actual metals today. According to a 2017 report by NewVantage Partners, 85% of companies desire to operate data-drivenly and by 2022, the global market for data science platforms is expected to grow from $19.75 billion to $128.21 billion.

Without practical applications, data science is merely an abstract idea. Many businesses, however, find it difficult to revamp their decision-making processes to incorporate data. The issue is not a lack of information.

Below are the Possible Data Science Project Risks

- Data theft
- Data privacy violence
- Going out of budget
- Improper analytical
- Low data quality

Mathematical and Statistical Skills for Data Science

Today's world has made data science a popular technology. You must improve your statistical and mathematical expertise to learn data science.

Math for Data Science

Every discipline has been impacted by mathematics. The extent to which

Notes

mathematics is used differs between fields. The following list includes some of the most typical forms of maths you may encounter in your data science career.

1. Linear Algebra: The development of machine learning algorithms depends critically on understanding how to construct linear equations. These are what you'll use to look at and study data sets. Loss functions, regularisation, covariance matrices and support vector machine classification are all applications of linear algebra in machine learning.
2. Calculus: The training of algorithms and gradient descent both employ multivariate calculus. You will learn about quadratic approximations, curvature, divergence and derivatives.
3. Statistics: When using classifications like logistic regression, discrimination analysis, hypothesis testing and distributions, machine learning relies heavily on this.
4. Probability: This is essential for testing hypotheses and for distributions like the probability density function and Gaussian distribution.

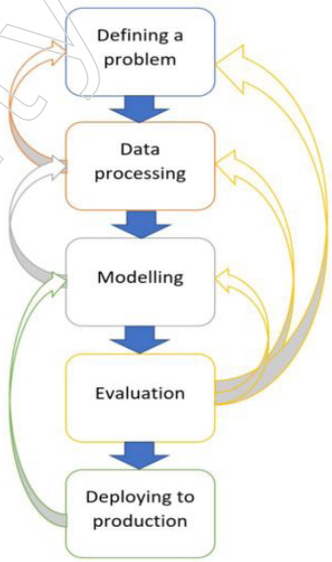
The study of data gathering analysis, visualisation and interpretation is known as statistics. A powerful sports car that operates on statistics is what data science is like. It processes raw data using statistics to produce the insights that go into the data products.

Additionally, by using statistics for data summarization and inference, you can gain a profound understanding of the data. In terms of these two terms, statistics is split into two categories–

1. Descriptive Statistics: To describe the data, descriptive statistics or summary statistics are utilised. It deals with analysing data quantitatively and summarising it. To summarise, graphs or numerical representations are used.
2. Inferential Statistics: Inferential statistics refers to the process of drawing conclusions or inferences from data. Can conduct numerous tests and draw inferences from the smaller sample using inferential statistics to draw a conclusion about the broader population.

Stages of Data Science Progress

In most cases, a Data Science project will have to go through five key stages: defining a problem, data processing, modelling, evaluation and deployment.



Notes

Stages of Data Science

1. Defining a problem: Any Data Science project must identify and specify a problem that has to be solved in its first stages. It might be challenging to determine how to approach a problem if it is not properly stated. This can include selecting the appropriate methodology for a Data Science project, such as classification, regression, or clustering.
2. Data processing: You can start working on the crucial chore of data processing after you know your problem, how you want to gauge progress and an idea of the methodologies you'll be employing. In any project including data science, this phase will typically take the longest and be the most crucial.
3. Modelling: The modelling portion of the Data Science project comes next and it's frequently the most enjoyable and interesting aspect. What the problem is and how you defined success in the first phase, as well as how you processed the data, will determine the structure this will take.
4. Evaluation: You must then know how to evaluate your models after they have been developed and put into use. This again refers to the problem formulation stage, where you will have determined your standard of success, but it is frequently one of the most crucial ones. If you don't evaluate your model properly, you can wind up with a model that is essentially useless and whose performance you are unsure of.
5. Deployment: Finally, you can put your model into production after a thorough evaluation and satisfaction with the findings. This might refer to several things, including whether you use the model's insights to change the way your organisation operates, if you use the model to determine whether changes you've made were successful, or whether the model is set up somewhere to continuously receive and assess real-time data.

Although the topic or subject matter of each Data Science project may vary, there are stages that are common to all Data Science initiatives. Problem definition, data processing, modelling, evaluation and deployment to production are all included in this. If one of these processes is missing from a project, it's likely that you'll draw the incorrect conclusions from it or that it was poorly designed.

Competency Required for Data Scientists

A cross-disciplinary set of abilities known as data science can be discovered when statistics, computer programming and domain knowledge meet. It consists of three separate yet related areas.:

- Statistics, to model and summarize data sets.
- Computer science, to design and use algorithms to store, process and visualize data.
- Domain expertise, necessary to formulate the right questions and to put the answers in context.

Other skills often missed are: –

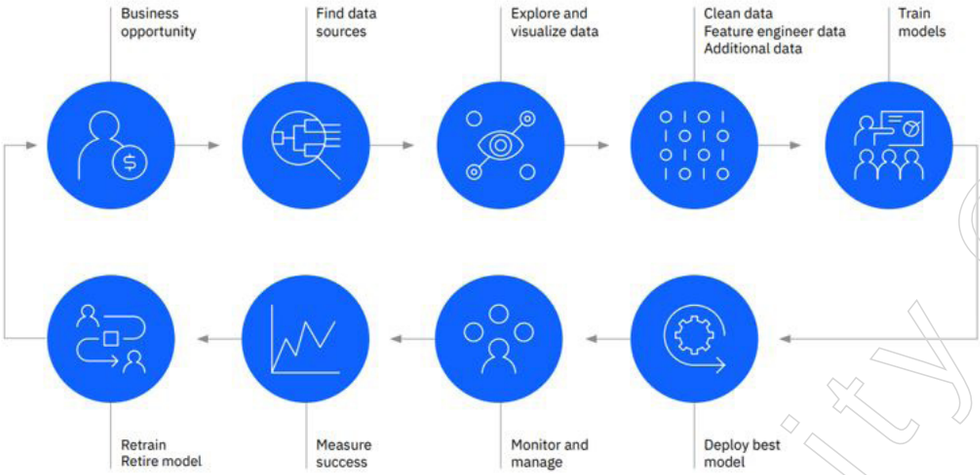
1. Leadership
2. Teamwork
3. Communication

The data scientist engages in or leads the AI enterprise workflow. They must:

1. Understand the business opportunity
2. Work with data engineers and the IT department to find the right data sources

Notes

3. Prepare data and build ML and specialized AI models
4. Assist in the deployment of models into the operations of the organisation
5. Measure success and communicate that back to the business



The first Data Science Apprenticeship Programme was established by IBM in the US in 2018. Director and Principal Leader for AI Skills Learning and Certification at IBM created a comprehensive blueprint outlining the competences needed for a data scientist as part of the programme. This design is suitable for:

1. Recruitment
2. kills development
3. Job expectations

Factors Influencing Data Science

As the world around us gets more digital and data-driven, the ability to work with vast amounts of data is becoming increasingly important. According to researchers, humanity will have produced 163 zettabytes of data by the year 2025. To be successful, an increasing number of companies are implementing data-driven strategies and utilising unique and cutting-edge techniques. Here are some elements influencing the use of data science in enterprises in the future:

1. Making data actionable for data science
2. Improving Operationalization
3. Accelerating 'time to value
4. A staggering amount of data growth
5. Shortage of data science talent

Predictive modelling has been utilised by businesses like Spotify, Netflix and Amazon to provide customers with the customised goods and services they want. In the future of analytics, companies will be able to make smarter decisions faster and with less effort than ever before because to new technology.

Issues and Challenges in Data Science

The actual issue is that organisations find it difficult to utilise the data they presently get in an insightful manner to produce knowledge that would aid in improved decision-making, risk management and threat defence.

Notes

Although this kind of "data fishing" does not adhere to the principles of efficient data science, it is nonetheless very common. Consequently, describing the issue clearly is the first step that needs to be taken.

The categories listed below are a few examples of how problems that data science can help solve can be categorised:

- Finding patterns in massive data sets: Which server in one's server farm requires the most upkeep?
- Detecting deviations from the norm in huge data sets: Does this assortment of purchases differ from what this customer has previously ordered?
- The process of estimating the possibility of something occurring: How likely is it that this person will choose to watch the video?
- illustrating the ways in which things are related to one another: What specifically is the topic of this internet article?
- Categorizing specific data points: Which animal do you believe this image represents, a mouse or a cat?

Types of Data Science Challenges/Problems

- Data Science Business Challenges: One of a data scientist's tasks while speaking with a line-of-business specialist about a business issue is to pay attention to key terms and phrases. Data science and artificial intelligence require a level of accuracy that must be recorded from the start:
  - ❖ Specify the problem that needs to be solved.
  - ❖ Be as specific as you can while answering each of the business questions.
  - ❖ Identify any additional business requirements, such as preserving existing client connections while maximising upselling and cross-selling opportunities.
  - ❖ Describe the anticipated benefits in terms of how they will impact the business, such as a 10% decrease in the rate of customer attrition among high-value clients.
- Real Life Data Science Problems: The application of hybrid mathematics and computer science models to current business problems to gain useful insights is known as data science. It is prepared to take the chance of dipping into the uncharted territory of "unstructured" data to gain important insights that help organisations improve their decision-making.
  - ❖ Managing the placement of digital ads through automated procedures.
  - ❖ The use of sophisticated analytics and data science will enhance the search function.
  - ❖ Making data-driven crime forecasts using data science.
  - ❖ Applying data science to prevent tax law violations.
- Data Science Challenges in Healthcare and Example: According to estimates, each person generates about 2 terabytes of data each day. These metrics include things like blood sugar, heart rate, tension and many others. To deal with such a huge amount of data, now have more advanced technologies, one of which is data science. This method assists in monitoring a patient's health by logging pertinent data.
- Data Science Problems in Retail: Although the term "customer analytics" is relatively new to the retail industry, the practise of using consumer data analysis

Notes

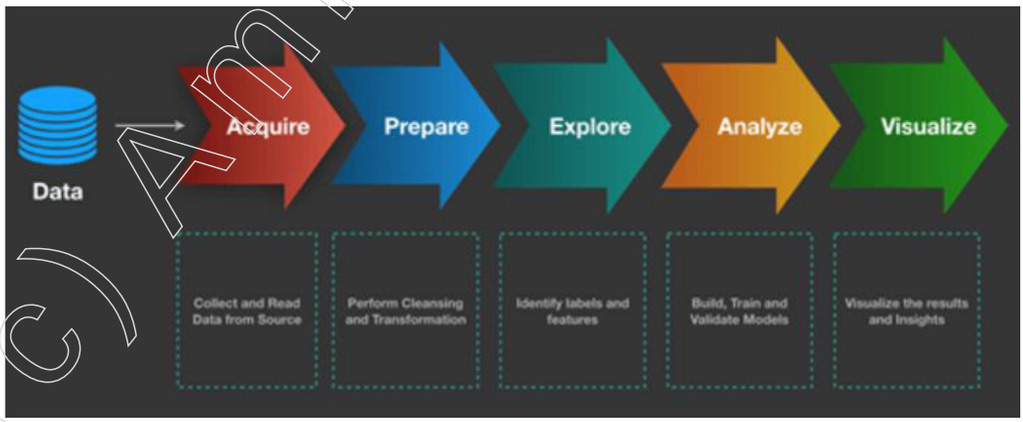
to offer customers customised goods and services is a centuries-old one. It is now straightforward to handle an increasing number of consumers thanks to the advancement of data science.

Real-time management of discounts and sales using data science tools may increase sales of previously discontinued goods and create hype for upcoming launches. Data science is also used to examine the entire social media ecosystem to predict which products will become popular soon so that they can be pushed to the market at the same time.

- Data Science Process: The following steps make up The Data Science Process, a systematic method for addressing data-related issues:
  - ❖ Problem Definition: Clearly stating the issue and the objective of the analysis.
  - ❖ Data Collection: Data collection and acquisition from numerous sources, including data processing and cleansing.
  - ❖ Data Exploration: Investigating the data to obtain knowledge and spot trends, patterns and connections.
  - ❖ Data Modelling: Creating algorithms and mathematical models to address issues and make predictions.
  - ❖ Evaluation: Utilising the right criteria to assess the model's performance and accuracy
  - ❖ Deployment: Using the model in a real-world setting to generate forecasts or automate decision-making.
  - ❖ Monitoring and Maintenance: In order to increase accuracy, the model's performance is being tracked over time and updated as necessary.
- Data Science Process: Most industries have realised how advantageous it is to harness the power of data with the introduction of Data Science and Machine Learning. This applies to all businesses, regardless of size, not just those in manufacturing, healthcare and automobiles. Having said that, the objective is to map the data to useful insights and then to money, regardless of the industry.

Every time data science is applied to a business, experiments are conducted first. These experiments go through multiple revisions before being prepared for production. The procedure, whether it be an experimental phase or a production phase, involves a straightforward series of steps that the data under examination is put through. Although the sequence may appear straightforward, the internal phases' intricacy can vary.

The following is a simple example of a Data Science Process Workflow:



Notes

Brief explanation of the data science workflow:

- ❖ Acquire: Getting the data for analysis is always the initial step in any workflow involving data science.
- ❖ Prepare: Obtaining the proper data is essential to any data science effort. Prior to conducting any analysis, you must collect the pertinent data, reformat it into a form that can be used for computing and clean it.
- ❖ Explore: Here is where you'll begin gaining high-level insights into what you're viewing and identifying the broad themes.
- ❖ Analyse: The fundamental part of data science is analysis, which involves creating, running and optimising computer programmes to examine and draw conclusions from the data gathered in the preceding stage.
- ❖ Visualise: Disseminating findings is the last stage of data science and it might take the shape of a data science product or written reports like internal memos, slide shows, business/policy white papers, or university research publications.

Benefits of Following Data Science Process

Over the past ten years, data science has revolutionised corporate growth. Amazingly, it is now possible to separate and organise certain predictive data to derive insights that are beneficial to your company. To boost the profitability of your company, you can also apply these insights in other fields like sales and marketing. However, this is not the end of data utilisation that is effective! There are numerous advantages to data science and justifications for using it in your company.

- ❖ Improves Business Predictions
- ❖ Business Intelligence
- ❖ Helps in Sales and Marketing
- ❖ Increases Information Security
- ❖ Complex Data Interpretation
- ❖ Helps in Making Decisions
- ❖ Automating Recruitment Processes

Challenges of Following Data Science Process

This study aids in determining the existing state of the company and potential areas for improvement. But comprehending data is not always simple. The obstacles that data scientists and data analysts encounter include data accumulation, security issues and a lack of the necessary tools.

- ❖ Data Quality and Availability
- ❖ Bias in Data and Algorithms
- ❖ Model Overfitting and Underfitting
- ❖ Model Interpretability
- ❖ Privacy and Ethical Considerations
- ❖ Technical Challenges

- Applications of Big Data

Large quantities of complicated, raw data are referred to as big data. By enabling data scientists, analytical modellers and other professionals to examine enormous

Notes

volumes of transactional data, today's businesses employ big data to make business more informed and enable business decisions. The precious and potent fuel that powers the enormous IT firms of the twenty-first century is known as big data. The usage of big data is becoming more widespread across all corporate sectors.

1. Travel and Tourism: Big Data is used by the tourism and travel industries. Can estimate the need for travel amenities at various places, increase sales through dynamic pricing and do so much more thanks to it.
2. Financial and banking sector: Banking institutions and consumer behaviour are aided by big data analytics, which consider buying habits, investment motivation and investment patterns.
3. Healthcare: Big data has started to fundamentally impact the healthcare industry with the help of predictive analytics, physicians, nurses and other healthcare professionals. It might lead to sole patients and personalised healthcare.
4. Telecommunication and media: Telecommunications and multimedia are the two sectors that employ big data the most frequently. Zettabytes of data are generated every day, making big data technology necessary.
5. Government and Military: The main industries using big data are telecommunications and the multimedia industry. Every day, zettabytes of data need to be processed, which calls for big data solutions.
6. E-commerce: A Big data application is e-commerce. It preserves consumer ties, which are crucial to the e-commerce sector. E-commerce websites use a variety of marketing techniques to attract customers to their stores, handle transactions and adopt improved methods of cutting-edge ideas to enhance enterprises using big data.

   Amazon: The fantastic e-commerce site Amazon receives a lot of everyday visitors. However, traffic on Amazon increases quickly during pre-announced sales, which could cause the website to crash. Therefore, it uses big data to manage this kind of traffic and data. Big Data assists in arranging and analysing data for long-term usage.
7. Social Media: The largest data generator is social media. According to statistics, social media, notably Facebook, generates 500+ terabytes of new data per day. Films, images, message exchanges and other visual media make up most of the information. A single social networking site action generates a significant amount of data, which is stored and examined as necessary. Terabytes (TB) of data are stored; processing them takes a long time. Big Data is a remedy for the issue.

Challenges of Big Data

Big data analytics is a complicated and difficult process and organisations encounter several difficulties when attempting to get insights and value from their data. Here are a few of the main difficulties faced by big data analytics:

- Data Complexity and Variety: Big data comes in a number of formats, including organised, semi-structured and unstructured data, making it challenging to manage and analyse.
- Data Quality: Big data is frequently unreliable, inconsistent, or imprecise, which might produce false insights and judgements.
- Data Security and Privacy: Big data must be safeguarded against unauthorised access and breaches since it frequently contains sensitive and private information.

Notes

- ❖ Scalability: The analytical infrastructure must be able to scale as data volumes increase to manage the increased load, which can be difficult and expensive.
- ❖ Talent Shortage: Data scientists and analysts with the expertise to successfully process and evaluate massive data are in limited supply.
- ❖ Integration: Integration of numerous systems and technologies is necessary for big data analytics, which can be difficult to establish and maintain.
- ❖ Data Governance: For big data to be managed and governed in a way that ensures compliance with rules and norms.
- ❖ Interpreting Results: Large and complicated datasets generated by big data analytics are frequently difficult to decipher and turn into useful information.

It takes a combination of people, processes and technology to tackle these problems. To fully utilise big data analytics, businesses must invest in strong analytical structures, data quality procedures, security and privacy guidelines and personnel development.

Big Data

Big data is the word used to explain the huge amount of organised and unstructured data that is produced and gathered daily by people, organisations and machines. Traditional data processing applications, which frequently have limits regarding their capacity to store, process and analyse massive datasets, cannot process data because it is too large and complicated.

Big data analytics involves gathering, evaluating and processing large amounts of data to gain knowledge and make wise business decisions. Big data's essential traits are typically summed up by the "3Vs": volume, velocity and variety.

Businesses need specialised equipment and software to manage big data, including the HDFS, MapReduce and YARN components of the Hadoop ecosystem, as well as Spark, HBase and Hive. Businesses must manage big data's technical hurdles as well as data privacy and security issues, as well as assure compliance with laws like the GDPR and CCPA.

How Big Data is Driving Digital Transformation

More data is being created and gathered right now than at any other time in the past. Social media networks are just one of the many sources from which this information is derived. from our phones and computers to wearable technology, scientific equipment, financial institutions and many other things.

Businesses now have the chance to comprehend consumers far more thoroughly than they previously could when using big data. Businesses increasingly get how important big data is to the digital revolution.

Companies that are undergoing digital transformation are better able to adapt to change and maintain their competitiveness in an increasingly digital world. The ability for an organisation to combine both in their efforts is where big data plays a part in digital transformation. facilitating corporate operations' digitization and automation as a result.

This automation and digitization are boosting innovation, increasing productivity and creating new business models.

Additionally, big data analytics is enabling organisations to monitor detailed data about certain or various client groups. This might be information on what they do when

Notes

they visit their websites, what they purchase, how frequently they do so and whether they plan to continue purchasing the same things.

As a result, businesses are making changes to better serve their consumers' future needs by utilising all this knowledge. developing objectives for how to fulfil these demands. The adoption of big data and data analytics is therefore necessary for firms to complete their digital transformation.

Some of the world's most brilliant minds started formulating theories about what would later become big data when the internet started to take off in the early 1990s. However, the phrase "Big Data" was originally used in 1999.

If you're still of the mindset that "Big data isn't required for my business," in 2020, you're missing out on a lot. Big data is now driving businesses and bringing about digital transformation; it is not only a catchy phrase that is gaining popularity. By combining business applications and productivity tools, the cloud can now give your company the knowledge it needs for a deeper understanding of its consumers.

Big data specialists have developed methods for using our glut of data to revolutionise not only how businesses run but also how they are managed. Big Data has amassed a sizable following in a short period of time for a reason. No matter your business's industry or size, data collecting, analysis and interpretation will have an impact on all businesses in several significant ways as they become more widely available.

Let's quickly go through a few examples of how Big Data affects different business sectors:

1. The Transformation
2. Manufacturing
3. Retail
4. Banking, Financial Services and Insurance (BFSI)
5. Media and Entertainment
6. In Closing

What is Hadoop?

Across dispersed networks, vast volumes of data are stored and analysed using the open-source Hadoop design. It was first created by Doug Cutting and Mike Cafarella in 2006 and the Apache Software Foundation currently looks after it. The Hadoop ecosystem is made up of several parts, including:

- Hadoop Distributed File System (HDFS): Data is stored on multiple cluster nodes using the distributed file system HDFS. Large files may be handled by it and because of its fault tolerance, data is always accessible even in the case of a hardware or software failure.
- MapReduce: Large data volumes can be processed across distributed systems using the programming model and software framework known as MapReduce. It is used to distribute data processing operations across several Hadoop cluster nodes in parallel.
- YARN: A programming language and software architecture called MapReduce is used to process huge data collections across distributed networks. It is used in a Hadoop cluster to divide data processing jobs among several nodes in parallel.

Notes

- HBase: Large amounts of structured data are kept in HBase, a NoSQL database. 230 Emerging Technologies for Business. It offers real-time access to data stored in HDFS and is built upon Hadoop.
- Pig: Hadoop uses the high-level programming language pig to process huge datasets. It offers a user-friendly interface for data processing operations and is made to make creating MapReduce jobs simpler.
- Hive: A data warehouse system called Hive is based on Hadoop. It offers a SQL-like interface for searching across big datasets kept in HDFS.
- Spark: Real-time and batch processing may both be handled by Spark, a quick and potent data processing engine. It offers a unified platform for data processing, machine learning and graph processing and is built on top of Hadoop.

As seen in the graphic below, the Hadoop ecosystem offers a complete and potent framework for the storage and processing of significant amounts of data across distributed computers. It is frequently used for data processing, analysis and machine learning in sectors like banking, healthcare, retail and telecommunications.

Data Scientists at New York University

and and and36 CREDITS | 2 YEARS FULL-TIME STUDY

There are two options to arrange the graduate program's 36-credit curriculum, giving students the chance to pursue a specialisation through the Industry Concentration or tracks. Within five years after enrolling at NYU as a candidate for the MS in Data Science, both part-time and full-time students are required to complete all 36 credits.

Scholarships

Selected applicants accepted into the programme will be given access to a limited number of tuition subsidies from NYU. On a competitive basis, all admission candidates will be taken into consideration for these awards. For up to two years, these scholarships will partially defray the cost of tuition.

Curriculum

The GSAS bulletin - Master of Data Science contains the degree requirements for the MS in Data Science.

The MS in Data Science (MSDS) degree programme consists of 36 credits. A capstone project that applies the theoretical information you learn in the programme to real-world situations is one of the program's essential components.

You will experience every step of fixing a real-world problem while working on the project, from gathering and processing real-world data to devising the most effective solution and eventually putting it into practise. The issues and data set you'll work with are drawn from actual contexts like those you may find in business, academia, or government.

The MSDS gives students the option to pursue the Industry Concentration, which enables them to use their degree program's knowledge and skills to apply to industry.

Students can choose a track to pursue via the MSDS. You can engage with academics through tracks who can provide guidance and assist you in creating a curriculum and professional goals.

Notes

The Master of Science in Data Science programme at NYU has a very competitive admissions process. This reflects both the growing popularity of the topic of data science and the exceptionally high grade of applicants to our programme.

Without exception, you must submit the following to support your application for admission:

1. GRE scores
2. TOEFL or IELTS; however, TOEFL is preferred (Required for all applicants whose native language is not English and who have not received a university degree in an English-speaking country)
3. Official college transcripts
4. Three letters of recommendation
5. Statement of Academic Purpose

For more information, visit the graduate school's application resource centre.

Educational Prerequisites

Successful MSDS candidates have undergraduate degrees in a wide range of fields, including statistics, computer science, mathematics, engineering, economics, business, biology, physics and psychology. The typical GPA for the 2022 entry cycle was 3.76.

Anticipate stronger grades in more pertinent subject matter from students coming from less selective colleges whose transcripts often only contain As and Bs. Regardless of academic standing, require particular and in-depth understanding of a few mathematical concepts as well as some background in programming and fundamental computer science.

You must have finished the following to be considered for the programme (or equivalents, e.g., MOOCs certification or course credit):

1. Calculus I: limits, derivatives, series, integrals, etc.
2. Linear Algebra
3. Intro to Computer Science (or an equivalent "CS-101" programming course): Have no set requirements as regards specific languages, but generally expect serious academic and/or professional experience with Python and/or R at a minimum.
4. One of Calculus II, Probability, Statistics, or an advanced physics, engineering, or econometrics course with heavy mathematical content

Preference is given to applicants who have significant more mathematical and/or computer science training than the necessary qualifications listed above, as well as those who have prior experience in operations research, data mining, computational statistics, machine learning and large-scale scientific computing (either in an academic or professional context).

Work Experience

The majority of our students contact us right out of college, but we also value evidence of prior employment with a specific employer and future work goals after the MSDS is complete. Business, government, academia and other fields might all be relevant depending on experience and desired employment domains.

Notes

Standardized Tests

Many of our students come to us right out of college, but also very much desire proof of pertinent work experience and definite employment goals after the MSDS is complete. Business, government, academia and other fields might all be relevant depending on experience and desired employment domains.

Please upload a PDF of the unofficial test results, which are made accessible once your test is over, to the "Additional Information Section" of your application in addition to delivering your official test results to the Graduate School of Arts and Science.

Want to be clear that there are no minimum GRE requirements and that when considering whether to admit a candidate, look at the entirety of their application. However, to give candidates a taste of things, the averages for the 2022 cohort of MSDS students are listed here:

- Average GRE Verbal: 160.0 (82 percentile)
- Average GRE Quantitative: 167.9 (91 percentile)
- Average GRE Analytical: 4.03 (56 percentile)
- Average TOEFL (where required): 110

It also demands proof of second-language English competency from some students who are required to submit it. It typically requires a TOEFL score of at least 100 overall (and much prefer greater scores) for such students and per university policies, it will not admit those scoring below that minimum.

Three Letters of Recommendation

Referees consistently hold applicants in the highest regard compared to other students or employees they have interacted with over the past few years and recommendations for accepted students are always good. The most weight is given to recommendations from academics or companies who can speak personally and in-depth about the applicant's situation, aptitude for and attitude towards data science projects. All correspondence should be on letterhead, while it's not necessary.

Internal Transfers/Current NYU Students

Internal transfers from other NYU graduate programmes are not permitted. Please be aware that you must submit a fresh application if you are a current graduate student at NYU and you wish to enrol in the NYU Data Science MS programme.

Ready to Apply?

It is advisable to start the application process if your background satisfies most of these criteria and you are motivated to provide the tools needed to fully utilise data. Please visit the Graduate School of Arts and Science website to apply.

Please examine the Graduate School of Arts and Science's general application policies page before beginning your application.

Deferral Requests

The Centre for Data Science will not grant a request to delay entrance. If an admitted student wants to postpone enrolling, they must decline the admission offer and reapply the following year. It will be necessary to use a different application entirely. At that later time, the new application will be evaluated with all other applicants.

Notes

What is the Difference: Neural Networks and Deep Learning

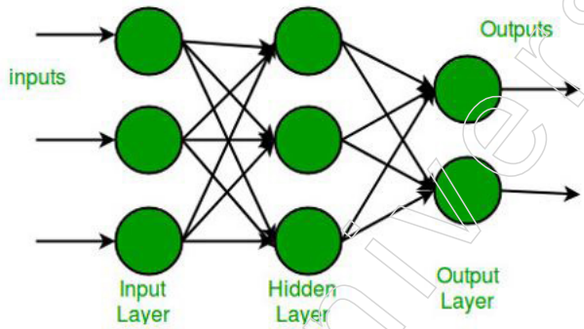
Machine learning and artificial intelligence have advanced significantly because of their inception in the late 1950s. In recent years, these technologies have become increasingly sophisticated and complex. Although technological developments in the field of data science are praiseworthy, they have led to a deluge of terminologies that are difficult for the average individual to understand.

There are many businesses of various kinds that employ these technologies, namely AI and ML, in their everyday operations. But many people struggle to differentiate between their several terms. Even the phrases “Deep Learning,” “Machine Learning,” and “Artificial Intelligence” are frequently used interchangeably.

Although there are numerous names for various concepts, most of them are closely related and have features in common, which is the cause of the misunderstanding. Nevertheless, every one of this terminology is distinct and helpful in its own way.

Before we discuss their distinctions, let’s talk about neural networks and deep learning systems separately.

What is a Neural Network?



Neural Network

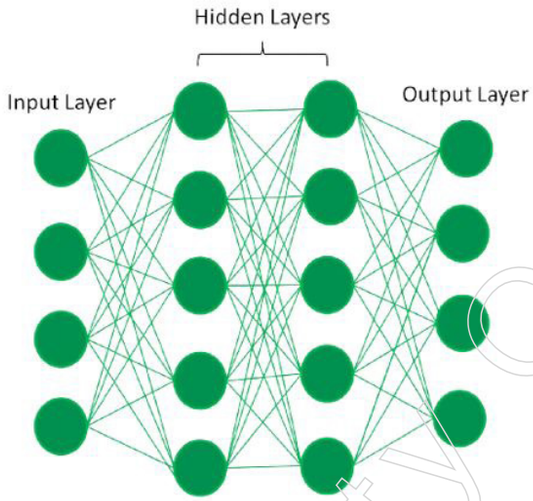
1. The human brain, the most complex object in the universe, serves as an inspiration for neural networks. Let’s start with comprehending how the brain functions. Neurons, which make up the human brain, are present. The simplest computational unit in every neural network, including the brain’s, is a neuron.
2. Until the processed output reaches the Output Layer, neurons receive information, process it and transmit it to other neurons located in the network’s many hidden levels.
3. Neural networks are algorithms that can classify or group the raw data and interpret sensory input using machine perception. All real-world data, including photos, audio, text, time series and other types of data, must be transformed. They are made to recognise numerical patterns found in vectors.
4. An Artificial Neural Network (ANN) only comprises three layers in its most basic configuration: an input layer, an output layer and a hidden layer.

What is Deep Learning?

In artificial intelligence, deep learning, often referred to as hierarchical learning, is a subset of machine learning that can imitate the computational power of the human brain and produce patterns resembling those employed by the brain to make decisions. Deep learning systems learn from data representations as opposed to task-based methods. It can gain knowledge from unstructured or unlabelled data.

Notes

What is a Deep Learning System?



Deep learning system

- A deep learning system, also known as a deep neural network, is a neural network with many of hidden layers and nodes in each one of those layers. Deep learning is the process of creating algorithms that may be used to train and predict outcomes from difficult data.
- A deep neural network, also known as a deep learning system, is a neural network having numerous hidden layers and numerous nodes in each hidden layer. Deep learning is the process of creating algorithms that may be applied to complex data to train and predict outcomes.

Architecture:

Neural Network architectures in detail:

- Feedforward Neural Networks – The input layer is on the top layer in this neural network architecture and the output layer is on the bottom layer. The middle layers are all concealed layers.
- Recurrent Neural Network – The connections between the nodes in this network architecture's ANNs form a directed graph along a temporal axis. As a result, over time, this kind of network displays dynamic behaviour.
- Symmetrically Connected Neural Networks – The sole distinction between them and recurrent neural networks is that symmetrically linked neural networks have connections between units that have the same weight in both directions.

Deep Learning model architectures in detail:

- Unsupervised Pre-Trained Network – This architecture is pre-trained based on prior experiences, as the name suggests and doesn't need formal training. Deep Belief networks and autoencoders are some of them.
- Convolutional Neural Network – This deep learning system can take an input image, give different items in the image meaning (learnable weights and biases) and tell these objects apart from one another.
- Recursive Neural Network – This is accomplished by continuously applying the same set of weights on a structured input and sending a topological structure rather than a scalar prediction on an input structure with changeable size.

Notes

Structure:

Neural Network structures in detail: A neural network has the following components

1. Neurons – A mathematical function called a neuron tries to imitate the actions of a biological neuron. It determines the weighted average of the supplied data before passing it through the logistic function, a nonlinear function.
2. Connections and weights – As the name imply, connections connect neurons in one layer to neurons in another layer or to neurons in a different layer. Each connection has a weight value assigned to it. A weight is used to indicate the strength of the link between the components. To reduce the likelihood of losing weight (error), the goal is to lower the weight number.
3. Propagation – There are two propagation functions in a neural network: backward propagation, which transmits the "error value," and forward propagation, which generates the "predicted value."
4. Learning Rate – Training neural networks involves the use of gradient descent. Backpropagation is used to determine the derivative of the loss function with respect to each weight value, which is subsequently deducted from that weight at each iteration. The learning rate controls how frequently and how slowly the model's weight values are changed.

Deep Learning model structures in detail:

A deep learning model has the following components

1. Motherboard – The deep learning model's motherboard chips typically rely on PCI-e lanes.
2. Processors – Based on the processor's core count and price, one must determine the GPU requirements for Deep Learning models.
3. Random Access Memory (RAM) – Deep learning models demand a significant amount of processing and storage. Larger RAMs are required for this.
4. Power Supply Unit (PSU) – It is crucial to have a large power supply unit that can handle massive and complex Deep Learning operations as memory requirements rise.

Due to their close association, it might be difficult to tell Deep Learning and Neural Networks apart on the surface. But you've undoubtedly already realised that deep learning and neural networks are two different concepts.

In contrast to neural networks, which rely on neurons to transmit data in the form of input to produce output with the help of various connections, deep learning is associated with the transformation and extraction of features that aim to establish a relationship between stimuli and associated neural responses present in the brain..

Machine Learning and Careers in DS

This branch of research tries to make machines more human-like in their actions and decision-making by giving them the ability to learn and develop their own programmes. There is very less human involvement in this; there is no explicit programming. Based on the machines' experiences during the process, the learning process is automated and improved.

Data science is not just a cutting-edge discipline that enables you to have a significant impact both inside your organisation and globally, but it is also one that

Notes

is expanding at a breathtaking rate. Big data and data science career prospects are blossoming as a rapidly growing number of organisations recognise the value of leveraging analytical data to improve business practises.

In fact, the Bureau of Labour Statistics (BLS) predicts that employment for statisticians in fields related to data science will increase 31% between 2021 and 2031, making this the fastest-growing vocation in the industry's mathematical sector.

Jobs in the data science field are anticipated to have good career prospects because there is a shortage of highly skilled individuals, according to numerous businesses. This is wonderful news for data science students and professionals since it signifies that there is a greater demand than there is supply for data scientists. Due to this lack, you'll discover that there are a wide variety of paths that a data science career might follow.

Although having options is usually a good thing, it may occasionally be challenging to comprehend how different professions differ and what sorts of skill sets and educational backgrounds are necessary for each. For those who are just getting started in the field of data science, this can be difficult.

Applications of Machine Learning

Machine learning is one of the most exciting technologies ever created. As the name says, it gives the computer the ability to learn, which makes it more resemble humans. Machine learning is currently being actively used in probably a lot more fields than one could imagine.

Machine learning is now being used by organisations to improve decision-making, increase production, discover diseases, forecast the weather and do many other tasks. Better tools are required to interpret the data already available and it is also necessary to prepare for the data that will be available given the exponential advancement of technology.

Some of the most common examples are:

- Image Recognition
- Speech Recognition
- Recommender Systems
- Fraud Detection
- Self-Driving Cars
- Medical Diagnosis
- Stock Market Trading
- Virtual Try On

1. How Data Science is Saving Lives

   Over the past two centuries, healthcare has made significant advancements. Live in an era of modern medicine, anaesthetic and preventative treatments, having transitioned from cleaning blood with leeches and biting sticks for pain alleviation.

   But there's still a long way to go.

   The burden of an ageing population, patients with many chronic diseases and underfunded infrastructures is felt by practitioners today. Making diagnostics, therapies and hospital operations more efficient and effective is the only way to solve these issues. And to achieve this, data science may be the answer.

Notes

Here are five ways that data science and analytics are improving healthcare today and paving the path for a brighter future for all of us.

2. Data-driven diagnosis:

   Early diagnosis is frequently essential for effective treatment. It's not always practical though, with doctors' offices already at capacity and hospital communications still dependent on fax machines. Sometimes things are overlooked, discovered after the fact, or never identified.

   The outcomes have led to patients being identified days, sometimes even weeks, faster and with improved accuracy, which is crucial given that diagnostic errors are responsible for up to 80,000 deaths annually in the US alone.

   This data-driven strategy is evident at Seattle Children's Hospital, where staff members have access to a comprehensive, on-demand view of crucial patient care data thanks to data from ten different source systems. Meanwhile, AI recently developed at University College London can identify cardiac problems in just four seconds.

3. Personalised medicine:

   The topic of customised healthcare, or "precision medicine," as it's commonly referred as in discussions about healthcare and data science, is always brought up.

   A one-size-fits-all approach to treatment is counterproductive because everyone has a unique biological make-up and upbringing. This is the basic tenet of precision medicine. Fortunately, technological advancements suggest that completely individualised treatments could not be too far off. All this boils down to how simple it is to examine someone's genetic makeup. Whereas it took 13 years to sequence the first human genome, can presently finish the job in a few hours. Additionally, the extraction of genetic information, which cost roughly £2 billion in 1990, may today be done for less than £200.

   Pharmaceutical companies can now begin demonstrating to insurance companies and regulatory authorities the benefit of more customised and consequently more successful therapies thanks to the technology's increased accessibility, which suggests that these treatments may soon be available on the market. Companies in the life sciences are also using real-world data and evidence to demonstrate the effectiveness of medications and novel therapies like gene therapy.

4. Improved hospital efficiency:

   Hospitals are busy settings and there is no space for error in an emergency. The repercussions of understaffing a ward or keeping patients waiting too long can be severe, which is why some hospitals have begun utilising data to ensure that this doesn't happen.

   Currently, four hospitals in Paris are collecting information from ten years' worth of admissions records to forecast demand peaks and using the knowledge gained to staff their wards appropriately. Another 40 hospitals will shortly copy it if it succeeds.

   Predictive analytics are also being utilised to optimise operating rooms at the University of Chicago Medical Centre. According to Forbes, these upgrades are anticipated to save the hospital $600,000 a year, a significant sum given the rising cost of healthcare. Of course, healthier patients and happier staff are the other major advantages.

5. Self-service healthcare:

   Today, making a doctor's appointment is not at all simple. The good news is that

Notes

you might stop needing one soon. Over the past four years, the use of mobile health applications has tripled and as sharing and data analysis get easier, they may eventually be able to diagnose health issues more quickly than a doctor's visit.

One pioneer in this field is UK-based Babylon Health, which created a healthcare app with the help of the NHS and Bupa and currently has 1.4 million users across Europe, Asia and Africa.

6. Faster drug discovery:

Bringing a new medication to market typically costs $2.6 billion and takes 12 years, according to Springboard. Big data analytics and machine learning can greatly speed up clinical research, which will drastically lower R&D costs.

Recent studies have used data to model the body's response to medications under various circumstances, leading to quicker regulatory agency approvals and more effective therapies.

A brighter healthcare future for all

There are countless ways that data science may benefit the healthcare sector and the bright people who work inside it, from diagnosis and treatment to hospital operations and aftercare. And it appears that analytics may contain many of the solutions that disadvantaged practitioners and institutions require.

Not all diseases are fatal. Some patients' bodies have a better chance of waging the battle against the illness. What recently increased these patients' resiliency and how might this be supported?

Like artificial intelligence, data science has its roots in the previous century and has just recently gained in popularity. This discipline enables the organised study of data as well as a wide range of tools and algorithms for change perception.

Thus, the increased level of acquired knowledge can shine light on information and circumstances that otherwise would be obscure to the human intellect. Even if they are not the same as predictions, patterns and future scenarios help us approach today's challenges with greater assurance.

Use cases in healthcare

Data analysis can be utilised in the medical field to provide more sophisticated and efficient treatments. hence lengthening the average population's lifespan. Many aspects of healthcare, including improved diagnosis, drug research and prevention, involve data science. Professionals must first collect the necessary data to complete the analysis before they can do this. Every day, the human brain and muscles generate two gigabytes of data. This amount of information is already sufficient to understand the body's secret processes.

Treating cancer:

Cancer, which is expected to cause 1.8 million diseases in the United States, of which more than 600,000 will be fatal, is still the number one public adversary.

BPM 31510 is a medication being developed by a Boston-based healthcare business that detects and initiates the natural death of diseased cells. These medications enable the body to naturally rid itself of cancer cells without the need for heavy medicine or any harm to the patient's health.

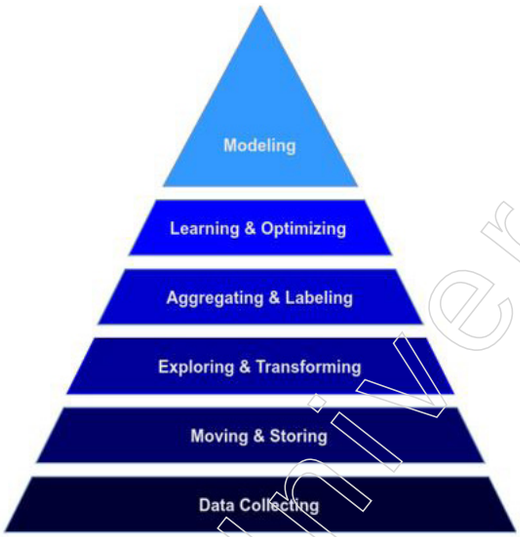
Notes

How Companies Should Get Started in Data Science

Many businesses are attempting to become data- and even AI-driven these days. Some of the most often used terminology in relation to the digital transformation of businesses are data science, machine learning and artificial intelligence (AI). One could believe that these technological advancements would address all corporate issues. In businesses, data scientists are frequently referred to as "digital wizards" who can easily transform data into useful insights and effective recommendation systems.

IT Infrastructure

Data Science requires a strong foundation before it can be utilised and does not offer solutions to all business problems. If it is not configured appropriately, the IT infrastructure within organisations might become a showstopper for Data Science projects. The graph below illustrates how data science is built on a hierarchy of needs:



Data Science Hierarchy

The base of the pyramid's needs, data collection, is created by software engineers. Naturally, the type of data collected depends on the company's operations and output. For a product that is intended for consumers, for instance, capture all data regarding how users engage with the product.

In this situation, directly retrieve the data from the programme that creates the user data. The data source may also be sensors put on equipment, structures, vehicles, or other objects from which data is collected.

Solutions for moving and storing the acquired data are needed for the following stage. As a result, the company's data engineers are required to ensure that the data can pass through the system with reliability. End-to-end data continuity is difficult to realise in practise and is a big problem if it is not set up correctly from the start.

A company has different options to store data:

On-Premises Data Storage

1. Servers owned and managed by the organisation itself
2. The greatest amount of control over network and data
3. Requires building and maintaining the IT infrastructure

Notes

4. Hardware: HDDs (good performance at reasonable costs), SSDs (highest performance, highest cost)

Hosted Data Storage:

1. Data storage kept off-site (at a data centre or cloud provider)
2. Cloud often cheaper than locally hosted options
3. Scalable, can be adapted to meet business requirements
4. Storage of sensitive data needs attention, data control is a concern and needs case-by-case solutions
5. Cloud providers such as AWS, GCP, or Azure provide a large stack of technologies in their ecosystem to turn your raw data into meaningful insights, products, or services

Hybrid Data Storage:

1. Use external providers for non-sensitive data
2. Store sensitive data locally

Companies will also need to make decisions on how to manage their data. Data warehouses are conventional methods for combining data from numerous sources. Business users may simply acquire insights and make decisions using the clean, consistent and performant model that data warehouses make.

A new position like the Data Scientist began to emerge as data became a highly asset, looking for value in the data. The ad hoc and agile nature of Data Exploration and Transformation that led to the development of Data Lakes is necessary for this new function.

One of the initial actions data scientists do to better comprehend the data is the exploration and transformation of the data. To prepare for the following steps in the Data Science process, it also involves wrangling the data. Because they can manage enormous volumes of data from several sources, utilise cutting-edge technology like Spark and enable novel approaches to data analysis like graph analytics, recommender systems and predictive analytics, data lakes have grown in popularity.

Data Culture

Businesses with a strong data culture frequently use data to inform their decisions. Data is viewed as the primary resource for gaining insights into each department of the organisation according to the idea of a data-driven culture. Providing staff with the tools to actively query and use data is the goal. Access to data, data governance, methodological expertise in data analysis and the necessary infrastructure, as previously noted, are all necessary for this.

The foundation of data science is data. You cannot use data science approaches without any data. The quantity and quality of the data are the two key factors consider when leveraging data science.

Amount of Data

The amount of information needed to conduct thorough analyses varies on the research objectives, the techniques used and the project's broader context. High quantities of data are typically required for advanced machine learning techniques to fit the models.

Notes

The training of machine learning models uses features. The quantity of information needed to generalise with accuracy will increase exponentially as the number of features increases. Must comprehend dimensionality and how it connects to algorithms to comprehend this.

Data Quality

Even more so than the quantity of data, the quality of the data determines how valid the results are. More rubbish simply produces a more accurate estimation of the incorrect object. Data scientists must constantly have a thorough understanding of the underlying data because of this. And if they don't (due to their inexperience in the company or their need to transition between other projects), they will need to collaborate closely with specialists who have a thorough comprehension of these data.

The reasons for dirty, low-quality data are manifold:

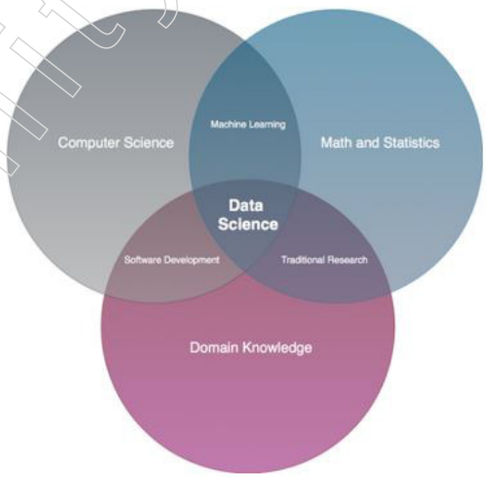
1. User entry errors
2. Legacy systems
3. Data migration
4. Poorly applied coding standards/programmer error
5. Evolving applications

Can investigate the following aspects to measure data quality:

1. Validity: complies with a schema
2. Accuracy: meets criteria for gold-standard data
3. Completeness: dataset contains all records
4. Consistency: the dataset is consistent with other datasets (from other sources)
5. Uniformity: values use the same units of measurement

Skilled Employees:

Data-driven businesses require personnel with advanced skills. Data science is a dynamic area that develops quickly. Technologies that are now considered de-facto standards, such as Spark for large-scale data processing, were not even invented ten years ago. Combining computer science, math, statistics and domain knowledge is necessary for data science.



Skilled for data scientist

Notes

Practitioners of data science are frequently obliged to express their key ideas and conclusions in non-technical terms so that a wider audience can comprehend them. Considering this, employers may also want to search for data scientists that have excellent communication abilities.

The combination of the skills is uncommon and businesses have trouble locating and keeping employees with these qualifications. Since Data Scientists are in such high demand across sectors, paying them is undoubtedly the first challenge. The average compensation for a Data Scientist in the US is $120,334 per year, according to the employment website Indeed.

The average income varies per country in Europe and is lower. The average annual wage for a data scientist in Germany, according to Glassdoor, is €60,000. The average salary in the UK is £53,628 and in France it is €45,198. Switzerland pays the highest salary in Europe, at about €88,400.

Corporate Strategy:

It's likely that AI and data science are not central to your organisation's strategy unless you work for a tech company in the Silicon Valley region. And given that not all businesses are expected to become leaders in the tech sector, this might be okay for some of them. However, these subjects and their enablers are vital for the future if your business is engaged in a digital transformation path with the goal of becoming a data-driven organisation.

The issue is that it is simple to write these buzzwords on a presentation and claim to be utilising all the new technologies, but it is much more difficult to gain a thorough understanding of what this means for the organisation, determine how AI helps to speed up processes and implement these technologies to eliminate time-consuming, repetitive tasks.

How Can Someone Become a Data Scientist

There is substantial disagreement over whether education is required to work as a data scientist. Although many experts have entered the field via alternative means, a university degree can undoubtedly be helpful.

Most graduate programmes and positions favour candidates with degrees in disciplines like computer science, data science, mathematics, statistics, engineering and even physics. Some programmes, however, will educate anyone with a degree in data science.

These are the technical skills you'll need to develop to become a fully-fledged data scientist:

- Python
- R
- Statistics and math
- SQL and NoSQL
- Data visualisation
- Machine learning
- Deep learning
- Natural language processing
- Big data

Notes

- Cloud computing

These are the cross-functional, non-job-specific talents that are now more commonly referred to as “power skills” or “human skills.” Those for a data scientist include:

- Communication
- Storytelling
- Critical thinking
- Business acumen
- Problem-solving
- Teamwork

These are some of the most frequently listed phases, however the precise needs for data scientists will vary depending on several circumstances:

- Learn data wrangling, data visualisation and reporting
- Work on your statistics, math and machine learning skills
- Learn to code
- Understand databases
- Learn to work with big data
- Get experience, practice and meet fellow data scientists
- Take an internship or apply for a job
- Follow and engage with the community

Recruiting for Data Science

A data scientist is skilled at interpreting data, which necessitates the use of statistical techniques and equipment but primarily relies on analytical thinking. These technical professionals are relied upon by companies across numerous industries to gather, clean and validate their data. Data scientists uncover patterns and use them to make data-driven decisions, improve corporate processes and adjust strategy. They do this with tenacity and software engineering abilities.

There are some common abilities that many recruiters and hiring managers look for in data scientist applicants, even though every organisation has its own unique needs. You may select the people who are the best fits by knowing which abilities and qualifications are necessary and which are desirable.

Required skills and qualifications	Preferred skills and qualifications
7+ years’ experience in data science	Master’s degree in stats, applied math, or related discipline
Proficiency with data mining, mathematics and statistical analysis	2+ years of project management experience
Advanced pattern recognition and predictive modelling experience	Professional certifications
Experience with Excel, PowerPoint, Tableau, SQL and programming languages (i.e. Java/Python, SAS)	
Comfort working in a dynamic, research-oriented group with several ongoing concurrent projects	

Notes

Careers in Data Science

Here are some of the most common job titles and careers in data science:

1. Business Intelligence Analyst: By studying data and creating a clearer picture of the company's position, ABI analysts use data to assist identify market and business trends.
2. Data Mining Engineer: The data mining engineer looks at information gathered from third parties as well as data from their own company. A data mining engineer will also develop sophisticated algorithms to aid in the analysis of the data.
3. Data Architect: Data management systems use blueprints that are developed by data architects in collaboration with users, system designers and developers to centralise, integrate, maintain and safeguard data sources.
4. Data Scientist: Data scientists start by converting a business case into an analytics agenda, creating hypotheses, comprehending data and exploring patterns to gauge their effects on organisations. Additionally, they look for and select algorithms to aid in additional data analysis. Business analytics is used by data scientists to explain what impact the data will have on a company in the future and to help the organisation come up with ways to deal with these consequences moving ahead.
5. Senior Data Scientist: A senior data scientist can predict what requirements a company will have in the future. In addition to acquiring data, they thoroughly examine it to effectively address extremely complicated business challenges. Their expertise allows them to not only establish new standards but also to spearhead their creation, as well as ways to employ statistical data and tools to aid in future data analysis.

   The position of a data scientist has emerged as the most in-demand career of the decade, with millions of job openings in the Big Data industry globally. Companies nowadays are employing data scientists' insights to stay one step ahead of the competition while minimising administrative costs in the data-based environment. There are frequently job openings for data scientists at well-known companies like Oracle, Apple, Microsoft, Booz Allen Hamilton, State Farm, Walmart and others. With the help of our data science interview questions, you can get through your interview on the first go.

   Businesses will require data scientists if they want to create successful, can't-miss strategies and make data-driven decisions.

High School Students and Data Science Careers

According to Business Insider, the top job in the United States for the past four years has been data scientist. Additionally, according to the U.S. Bureau of Labour Statistics, through 2026, the field of data science will experience a far greater increase in employment than the national average for all occupations, at a rate of 27.9 percent. High school students should be given the chance to prepare for the diverse range of data science occupations because of this. The 10 industries listed below may motivate your students to learn more.

Communications, Media and Entertainment

Instant entertainment is always available to us. Customers are accustomed to being able to get what they want when they want it. Data science is used by businesses like YouTube, Netflix and Spotify to provide the most relevant content to their audiences.

Notes

Digital Marketing

Data science is necessary for digital marketing to gather information and make business decisions. Marketing managers can predict user behaviour and anticipate customers' demands with the aid of data-driven information. Companies like HubSpot, GetResponse and Signpost use data science to personalise offers and identify patterns and trends.

Gaming

The idea of converting a favourite hobby into a job opportunity is the dream of many teenagers. The global gaming market generates $180 billion a year in revenue with over 3 billion players. To improve game models, data scientists who work in the gaming industry construct models, examine optimisation points, make forecasts and find trends. Additionally, they evaluate consumer habits generally to boost business profitability and catch fraud. The development of machine learning algorithms enables the quicker detection of suspect account activity.

Retail

In the highly competitive world of retail, organisations get an edge by being able to predict and satisfy client needs. Data scientists offer insights that keep customers of merchants satisfied and coming back to their establishments. Consumer happiness is ensured by using data to provide tailored and relevant purchasing experiences, which encourages repeat business.

Healthcare

Data scientists are in high demand in the healthcare sector, which is increasing quickly. Doctors and scientists have a vast amount of data on which to base their studies, with the United States producing over 1.2 billion clinical documents each year. Being able to collect, structure and process this data helps researchers gain a deeper understanding of the human body. Furthermore, wearable trackers that monitor heart rate, sleep patterns and brain activity assist medical professionals in managing patient care.

Finance

One of the earliest sectors to apply data science was the financial sector. Data scientists assist banks with a variety of tasks, including fraud detection, customer analysis and experience and risk assessment and monitoring. Making decisions based on facts allows for a more stable financial climate.

Cyber Security

According to Cybersecurity Ventures, worldwide cybercrime will rise 15% annually by 2025, when it would total a staggering $10.5 trillion annually (up from $3 trillion just six years ago). By analysing attack patterns and coming up with ways to defend against them and stop them from happening again, data scientists can utilise artificial intelligence and machine learning techniques to locate the origin of hostile actions.

Government

Data scientists play a crucial role in a variety of public service sectors, including fraud detection, energy exploration and financial market monitoring. Data science has evolved into a potent tool for battling the COVID-19 pandemic and perhaps any other future infectious disease epidemic.

Notes

Travel

Travel and tourism are now more accessible than ever to a wider range of people. Data scientists are essential to the effective and profitable operation of the tourism industry. Through personalised marketing, brand development and trip planning optimisation, consumer behaviour analysis assists airlines, hotels and booking websites in continually improving their services.

Education

According to Stanford University academics, the field of education lends itself particularly well to data science. To manage and enhance student learning and teaching, district records, digital archives of instructional materials and grade books, as well as the mountain of student data that classroom instructors collect each day, all provide valuable information.

All this data can be mined and examined to comprehend and address enduring educational issues, support the development of new teachers and improve training. Additionally, data scientists can assist online learning platforms in optimising their course content for students.

Case Study

The most crucial part of any data science interview is to demonstrate your ability to apply your knowledge to practical use cases. Study the data science case study below to learn how to analyse and resolve a problem. The following data science case studies are resolved and discussed in Python.

Case Study on Text Emotions Detection

This use case is for you if you're one of those people interested in natural language processing. A machine learning model will be trained to create emojis from text input. So, artificially intelligent chatbots can be trained using this machine learning methodology.

Use Case: The face, gestures, speech and written words are just a few of the many ways that people can convey their feelings. Text emotion detection is an issue in content-based categorization. Since a person can communicate his or her emotions in a variety of ways, it can be challenging to identify those feelings in text written by a person.

In applications like chatbots, customer service forums, customer reviews, etc., identifying this type of emotion from a text authored by a human is crucial. To recognise the mood of a text, you must train a machine learning model that can do so by showing the most appropriate emoji based on the input text.

```
This looks so impressive 😳
I have a fear of dogs 😱
My dog died yesterday 😢
I don't love you anymore..! 😂
```

The output of Text Emotions Detection

Summary

- The decision-making processes of many organisations may be supported by the real-time processing of this data. Data science is a field that focuses on the processes of gathering, integrating and processing massive amounts of data to provide information that can be used to make informed decisions.

Notes

- Data science is a branch of study that combines domain knowledge, coding abilities and math and statistical knowledge to derive practical insights from data.
- The roles in computer science vary depending on the skills and interests of a candidate, but most positions in this field enhance technology and create or maintain digital products for a business. Some occupations in computer science also include the physical components and hardware of a computer system.
- The route to become a data scientist involves moving up the ranks from entry-level analytic positions and assuming increasing responsibility and leadership roles. To make better business decisions and address complicated issues, data scientists combine data analysis and business intelligence
- A career as a data scientist promises a wealth of potential and excellent compensation because they are experts in statistics, data science, big data, R programming, Python and SAS.
- A data science project will often need to go through five crucial stages: problem definition, data processing, modelling, evaluation and deployment.
- Predictive modelling has been utilised by businesses like Spotify, Netflix and Amazon to provide customers with the customised goods and services they want.
- With the help of new technology, businesses will be able to make better decisions faster and with less effort than ever before in the future of analytics. Data scientists will be in more demand as businesses adopt automation.
- The Data Science Process is a systematic approach to solving data-related problems.
- The precious and potent fuel that powers the enormous IT firms of the twenty-first century is known as big data.
- Big data analytics is a complicated and difficult process and organisations encounter several difficulties when attempting to get insights and value from their data.
- To fully utilise big data analytics, businesses must invest in strong analytical structures, data quality procedures, security and privacy guidelines and personnel development.
- Machine learning is now being used by businesses to enhance decision-making, boost production, find diseases, predict the weather and perform many other things.
- This data-driven strategy is evident at Seattle Children's Hospital, where staff members have access to a comprehensive, on-demand view of crucial patient care data thanks to data from ten different source systems.
- The foundation of data science is data. You cannot use data science approaches without any data. The quantity and quality of the data are the two key factors must consider when leveraging data science.

Glossary

- Machine Learning: In the field of artificial intelligence known as machine learning, a system is programmed to carry out a certain task automatically.
- Algorithms: Algorithms are a particular collection of guidelines or procedures that are applied in a computation to address issues or complete a task.
- Statistical models: Mathematical models known as statistical models describe the connections between random and non-random variables.

Notes

- Regression Analysis: To generate a real number value that represents a quantity on a line, regression analysis is a statistical technique that assesses associations between a dependent variable and independent variables. Temperature, sales turnover, for instance.
- Programming: Computer programming languages are used to design and build the models that are utilised in data analysis. Programming can also be used to organise data, clean data and help visualise data so that stakeholders can understand it.
- Statistics: The study of data gathering analysis, visualisation and interpretation is known as statistics.
- Descriptive Statistics: To describe the data, descriptive statistics or summary statistics are utilised. It deals with analysing data quantitatively and summarising it. To summarise, graphs or numerical representations are used.
- Inferential Statistics: Inferential statistics refers to the process of drawing conclusions or inferences from data.
- Big Data: Large quantities of complicated, raw data are referred to as big data.
- Neural Network: The neurons of a neural network are modelled after those in the human brain. It is composed of several neurons that are intricately coupled to one another.
- Deep learning system: Deep learning neural networks are distinguished from neural networks based on their depth or number of hidden layers.

Check your Understanding

1. What is/are the area/s in which data science is useful?
   a) Making decision
   b) Developing more accurate projection
   c) Find pattern inn different data set that are similar
   d) All the above
2. _____ are a particular collection of guidelines or procedures that are applied in a computation to address issues or complete a task.
   a) Algorithms b) Machine Learning
   c) Statistical Models d) None of the above
3. What is/are the role/s of data science?
   a) Data Architect b) Database administration
   c) Statistician d) All the above
4. What is the job in computer science?
   a) Data Analyst b) Data Scientist
   c) IT Analyst d) All the above
5. What are the job titles in computer science
   a) Data engineer b) Web developer
   c) Business analyst d) Data scientist
6. What are the possible Data Science project risk?
   a) Proper analytical b) High data quality

Notes

c) Data Theft d) Data privacy

7. How much stages do a Data Science progress has?

a) 6 b) 5

c) 4 d) 7

8. Which business utilise predictive modelling?

a) Spotify b) Netflix

c) Amazon d) All the above

9. How many steps are there in Data Science process?

a) 7 b) 5

c) 8 d) 4

10. What are the benefits of Data Science process?

a) Improve business predictions b) Helps in sales and marketing

c) Complex data interpretation d) All the above

11. The precious and potent fuel that powers the enormous IT firms of the twenty-first century is known as _____.

a) Data Science b) Big Data

c) Machine Learning d) None of the above

12. A Big data application is _____.

a) Travel and tourism b) Healthcare

c) E-commerce d) Amazon

13. Which is the largest data generator?

a) Social Media b) E-commerce

c) Healthcare d) Banking sector

14. What are the challenges of big data?

a) Data Quality b) Scalability

c) Integration d) All the above

15. Within five years after enrolling at NYU as a candidate for the MS in Data Science, both part-time and full-time students are required to complete all _____.

a) 28 Credits b) 36 Credits

c) 32 Credits d) 30 Credits

16. Without exception, what must be submitted to support application for admission in NYU.

a) GRE scores b) TOEFL

c) IELTS d) All the above

17. _____ is the part of neural network structure.

a) Learning rate b) PSU

c) RAM d) Processor

18. The base of the pyramid's needs, data collection, is created by ____.

a) Software engineer b) Data Scientist

Notes

c) Data Specialist d) None of the above

19. How many ways a company can store data?

a) 4 b) 3

c) 2 d) 5

20. What is/are the skill required for to be a data scientist?

a) Python b) R

c) Data Visualisation d) All the above

Exercise

1. Briefly explain the basic concept of Data Science.
2. Explain the difference between computer science and Data Science.
3. Explain how Data Science is useful.
4. What are the different stages of Data Science?
5. What are the challenges of Data Science?
6. Explain the process of Data Science.
7. Explain the concept of big data.
8. What is the utilisation of machine learning?
9. What is the different type of opportunity in Data Science?

Learning Activities

1. Explain with example how Data Science is helpful in medical science.
2. Describe different job opportunity in data science.

Check your Understanding-Answers

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

Further Readings and Bibliography

1. Field Cady. The Data Science. 2017
2. William Vance. Data Science: 3 Book in 1 – Beginner's Guide to learn the Realm of Data Science. 2020
3. Peter Bruce Andrew Bruce. Practical Statistics for Data Scientist. 2020
4. Reema Thareja. Data Science and Machine Learning using Python. 2022
5. Uma Maheshwari R Sujatha. Introduction to Data Science: Practical Approach with R and Python. 2021

Module - II: Introduction to "R"

Learning Objectives

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

- Understand the R-basics
- Explain differentiation of R and S programming
- Discuss installation of R
- Describe R add inns
- Explain how to import data from different file formats
- Summarise concept of R-controls
- Understand R – functions
- Know different types of loop function
- Learn different types of R tool

Introduction

R is a comprehensive collection of tools for math, data manipulation and graphic display. It has, among other things

- a facility for managing and storing data efficiently,
- a collection of operators for array calculations, particularly matrices,
- a substantial, cogent and comprehensive collection of intermediate data analysis tools,
- graphical tools for data processing and display on a computer screen or in hardcopy and
- Conditionals, loops, user-defined recursive functions, input and output facilities and a well-developed, straightforward and efficient programming language (named "S") are all included in the language. (In fact, the majority of system-supplied functions are also written in the S language.

R is a major platform for recently created interactive data analysis techniques. It has expanded quickly and is supported by a sizable number of packages. However, most R programmes are essentially transient and created for a single data analysis task.

2.1 R- Basics

R is frequently used to study and visualise data through graphical presentation and statistical computing. R is a popular computer language used for graphical display and statistical computation. It is most frequently used to examine and display data.

Notes

Why Use R?

- Excellent source for data science, machine learning, data analysis and visualisation
- Offers a variety of statistical methods, including data reduction, classification, grouping and statistical tests.
- Simple to create graphs like box plots, scatter plots, histograms, pie charts, etc.
- Functions across various platforms (Windows, Mac, Linux)
- Open-source and free
- Enjoys widespread community backing.
- Has a variety of packages (function libraries) that can be utilised to address a variety of issues.

R is a programming language that two statisticians at the University of Auckland in New Zealand originated and developed in 1991. It wasn't until 1995 that it was formally made free and open source. It offers time series approaches, statistical and graphical techniques, linear and non-linear models and many other functions for its beginnings. Even though Python is the most popular language in the field of data science, R is still frequently used for niche applications, such as in financial services, research and healthcare.

2.1.1 What is R Programming

Low-level language R was developed to make it simple to implement the S language. As part of the GNU project, Ross Ihaka and Robert Gentleman developed this language in 1995. It was later made public in 1997.

The R foundation, a nonprofit organisation, has since taken up maintenance of this project, which made use of the GNU operating system and its packages to produce free software. It is therefore free and operates under the GNU General Public Licence, which is distributed in source code form by the Free Software Foundation.

Low-level languages like C and C++ can relate to R to carry out computationally intensive processes. Even C code can be written by users to modify R's objects. R has a very clear understanding of the context in which several functions exist and are made available by libraries and modules. R is frequently more effective and quicker than many other languages because these libraries are frequently created in a language like C while the user does the code in R.

You must be aware of the R programming language's background and place in the data science community to comprehend what it is. Statistical computing or running statistical tests on data to create statistical models, is the main task that R is used for.

Since R is an object-oriented language, all its operations revolve around objects. Anything that may be stored in a variable, such as one-dimensional data structures, two-dimensional data structures, user-defined functions, etc., can be considered one of these items. Considering that R is a low-level language, learning it is not too difficult. Another justification for the lengthier codes is this. R is comparatively one step above when compared to other low-level languages like C, nonetheless. Additionally, R is a dynamically typed language, which eliminates the need for variable declarations. R recognises the class automatically, which makes coding in R simple.

R is a large programming language and there are several things to consider if one wants to fully comprehend it. Several basic concepts of R language include-

Notes

1. Objects and Environment

In the Environment window, each of these things is displayed. Because the user can see the items that are now taking up RAM space, managing objects is made incredibly simple. To improve the efficiency of the coding process, the user can export, import and even remove objects using a R Data file. Common functions include:

Function	Formula
Finding all the names of objects in the environment	>> ls()
Saving all the objects that are there in the environment	>> save.image("MyBackup.RData")
Saving one object from the environment	>> save(Cities, file="cityobj.RData")
Removing an object	>> rm(Cities)
Removing all the object from the environment	>> rm(list=ls())
Loading a RData file	>> load("MyBackup.RData")

2. Console:

It is regarded as the brains behind the R programming language and R Studio, the IDE used to execute R. You may create code, run it and view the results all at once on the console. However, any code entered in the console cannot be saved as a script.

Additionally, console codes that have already been run cannot be changed. It's interesting to note that just the console is used to run any code that is entered into the code window. Get a handle on the R programming language's console notion by watching the James Cook video below.

3. Script

R files can be created from the codes entered in the code box. Commonly referred to as a R script, this file. These scripts make it easier to reuse and share the codes.

4. Operators

These are the symbols that let us carry out specific operations. For instance, the task completed with the help of the function sum() can also be completed with the help of its operator, like this:

sum(10,20) can also be performed using the + operator -> 10+20

In R there are several operators such as:

Type	Operators
Assignment Operators	= -> <-
Arithmetic Operators	/ (division) * (multiplication) + (addition) – (substraction) %% (modulus) Remainder %/% (Integer Division) Quotient ^ (Power)

Notes

Relational Operators	> (Greater than) < (Less than) <= (Greater than equal to) >= (Less than equal to) == (Equal to Equal to)
Logical Operators	& (AND) | (OR) For changing the data types, a process also known as type casting requires functions such as as.character(), as.factor(), as.numeric() etc. However, before changing the data types one must be aware of the hierarchy of typecasting where the highest data type is character and lowest is logical and a higher data type cannot be converted to a lower data type apart from a few specific exceptions.

For changing the data types, a process also known as type casting requires functions such as as.character(), as.factor(), as.numeric() etc. However, before changing the data types one must be aware of the hierarchy of typecasting where the highest data type is character and lowest is logical and a higher data type cannot be converted to a lower data type apart from a few specific exceptions.

2.1.2 What is the Difference Between R and S Programming

An environment and programming language for statistics and statistical graphics is called S. S-Plus and R are examples of S implementations. This indicates that both items are constructed from the same materials and carry out the same tasks (macros). Most frequently used statistical functions written in S will operate in R and vice versa.

S is a language that was created as part of a project at Bell Labs. By converting macros into functions, the development was made for data analysis, statistical modelling, simulation and graphics. One of the earliest statistical computer systems developed was the S language and since then, copies have been made; one such clone is the R language.

In contrast to the S-plus language, the most recent version is a free open-source programme that supports a large user base. The R language, which is a dialect distinct from the S language, has replaced the S language as the more common tongue.

The R language is like a primer for the S language, with the exception that some functions cannot be swapped out and must be modified. The usage of lexical scoping in R further simplifies coding by providing an object with a reference to arguments, variables, constants, types and functions.

The S language uses dynamical scoping to create global identifiers since it is more advanced. S-plus offers a more sophisticated user interface for graphics than R, which has a more basic user interface.

There are however several differences you should be aware of.

- S-Plus includes a highly developed graphical user interface (GUI), which differs greatly from R, in addition to typing in functions and other information. The graphical user interface for R is simple. (However, various initiatives to incorporate a GUI into it are in work.
- There are techniques to transfer data simply, however the standard data files for S and R are incompatible.

Notes

- The products' graphical output varies frequently since they each employ a different graphical engine and programming paradigm, especially when writing sophisticated visuals or/and including interaction.
- Completely distinct interfaces exist for other software, particularly databases.

Most operations that don't use visuals or simple graphics work without issue in both settings. However, if you use more sophisticated graphical tools, expect issues. Since this class focuses mostly on visual tools, there will be numerous discrepancies; less so in the tools themselves, but in the functions, which will not be compatible and have distinct names and implementations. These sections generally focus on R; significant differences with S are indicated and certain publications will be tailored specifically for S.

2.1.3 Functions for Reading and Writing Data

The R program's activities are now carried out on a terminal or prompt that is not saved anywhere. However, numerous programmes are developed in the software industry to save the data they get. One such way is to keep the information that was retrieved in a file. the following are the two most frequent operations that can be carried out on a file:

- ❖ Importing/Reading Files in R
- ❖ Exporting/Writing Files in R

Reading Files in R Programming Language

The complete data is lost when a programme is terminated. Even if the programme is terminated, storing in a file will protect our data. If there are many, it will take a long time to enter all the required information. However, by using a few commands, you may easily get a file's contents in R if it already has all the data.

Without making any changes, you can transfer your data from one machine to another with ease. Therefore, the files can be kept in a variety of forms. It could be kept in a.txt (tab-separated value) file, a.csv (comma-separated value) file, on the cloud, or in a tabular format. R offers much simpler ways to read those files.

File reading in R

Text files are one of the crucial file storage formats. There are several ways to read data from a text file in R.

read.delim(): This approach is utilised to read "tab-separated value" (or ".txt") files. Points (".") are used as decimal points by default.

Syntax: read.delim(file, header = TRUE, sep = "\t", dec = ".", …)

Parameters:

- ❖ file: the path to the file containing the data to be read into R.
- ❖ header: a logical value. If TRUE, read.delim() assumes that your file has a header row, so row 1 is the name of each column. If that's not the case, you can add the argument header = FALSE.
- ❖ sep: the field separator character. "\t" is used for a tab-delimited file.
- ❖ dec: the character used in the file for decimal points.

Example:

Notes

❖ R

```
# R program reading a text file

 # Read a text file using read.delim()

myData = read.delim("geeksforgeeks.txt", header = FALSE)

print(myData)
```

Output:

A computer science portal for geeks.

Note: The above R code, assumes that the file "geeksforgeeks.txt" is in your current working directory. To know your current working directory, type the function getwd() in R console.

read.delim2(): This method is used for reading "tab-separated value" files (".txt"). By default, point (",") is used as decimal points.

Syntax: read.delim2(file, header = TRUE, sep = "\t", dec = ",", …)

Parameters:

- ❖ file: the path to the file containing the data to be read into R.
- ❖ header: a logical value. If TRUE, read.delim2() assumes that your file has a header row, so row 1 is the name of each column. If that's not the case, you can add the argument header = FALSE.
- ❖ sep: the field separator character. "\t" is used for a tab-delimited file.
- ❖ dec: the character used in the file for decimal points.

Example:

Output:

1 A computer science portal for geeks.

file.choose(): In R it's also possible to choose a file interactively using the function file.choose() and if you're a beginner in R programming then this method is very useful for you.

Example:

❖ R

```
# R program reading a text file using file.choose()

 myFile = read.delim(file.choose(), header = FALSE)

# If you use the code above in RStudio

# you will be asked to choose a file

print(myFile)
```

Output:

1 A computer science portal for geeks.

- ❖ read_tsv(): This method is also used for to read a tab separated ("\t") values by using the help of readr package.

Syntax: read_tsv(file, col_names = TRUE)

Notes

Parameters:

- file: the path to the file containing the data to be read into R.
- col_names: Either TRUE, FALSE, or a character vector specifying column names. If TRUE, the first row of the input will be used as the column names.

Example:

- R

```
# R program to read text file
# using readr package
# Import the readr library
library(readr)
 # Use read_tsv() to read text file
myData = read_tsv("geeksforgeeks.txt", col_names = FALSE)
print(myData)
```

Output:

```
# A tibble: 1 x 1
  X1
1 A computer science portal for geeks.
```

Note: You can also use file.choose() with read_tsv() just like before.

```
# Read a txt file
myData <- read_tsv(file.choose())
```

Reading one line at a time

read_lines(): This method is used for the reading line of your own choice whether it's one or two or ten lines at a time. To use this method, reader package must be imported.

Syntax: read_lines(file, skip = 0, n_max = -1L)

Parameters:

- file: file path
- skip: Number of lines to skip before reading data
- n_max: Numbers of lines to read. If n is -1, all lines in the file will be read.

Example:

- R

```
# R program to read one line at a time
 # Import the readr library
library(readr)
 # read_lines() to read one line at a time
myData = read_lines("geeksforgeeks.txt", n_max = 1)
print(myData)
```

Notes

read_lines() to read two line at a time

myData = read_lines(“geeksforgeeks.txt”, n_max = 2)

print(myData)

Output:

[1] “A computer science portal for geeks.”

[1] “A computer science portal for geeks.”

[2] “Geeksforgeeks is founded by Sandeep Jain Sir.”

Reading the whole file

read_file(): The entire file can be read using this technique. must import the reader package to use this method.

Syntax: read_lines(file)

file: the file path

Example:

❖ R

R program to read the whole file

Import the readr library

library(readr)

read_file() to read the whole file

myData = read_file(“geeksforgeeks.txt”)

print(myData)

Output:

❖ A computer science portal for geeks.\r\nGeeksforgeeks is founded by Sandeep Jain Sir.\r\nI am an intern at this amazing platform.”

Reading a file in a table format

A tabular format is another widely used way to store a file. There are several ways to read data from a tabular data file using R.

read.table(): read.table() reads a file in table format using a general function. Data frames will be used to import the data.

Syntax: read.table(file, header = FALSE, sep = “”, dec = “.”)

Parameters:

❖ file: the path to the file containing the data to be imported into R.

❖ header: logical value. If TRUE, read.table() assumes that your file has a header row, so row 1 is the name of each column. If that’s not the case, you can add the argument header = FALSE.

❖ sep: the field separator character

❖ dec: the character used in the file for decimal points.

Example:

Notes

❖ R

```
# R program to read a file in table format
 # Using read.table()
myData = read.table("basic.csv")
print(myData)
```

Output:

```
 1 Name,Age,Qualification,Address
 2 Amiya,18,MCA,BBS
 3 Niru,23,Msc,BLS
 4 Debi,23,BCA,SBP
 5 Biku,56,ISC,JJP
```

read.csv(): read.csv() is employed to read "comma separated value" (.csv) files. The data will also be imported in this as a data frame.

Syntax: read.csv(file, header = TRUE, sep = ",", dec = ".", …)

Parameters:

- ❖ file: the path to the file containing the data to be imported into R.
- ❖ header: logical value. If TRUE, read.csv() assumes that your file has a header row, so row 1 is the name of each column. If that's not the case, you can add the argument header = FALSE.
- ❖ sep: the field separator character
- ❖ dec: the character used in the file for decimal points.

Example:

❖ R

```
# R program to read a file in table format
 # Using read.csv()
myData = read.csv("basic.csv")
print(myData)
```

Output:

```
Name Age Qualification Address
1 Amiya  18         MCA     BBS
2  Niru  23         Msc     BLS
3  Debi  23         BCA     SBP
4  Biku  56         ISC     JJP
```

read.csv2(): read.csv() is a variation used in nations where the comma "," serves as the decimal point and the semicolon ";" serves as the field separator.

Syntax: read.csv2(file, header = TRUE, sep = ";", dec = ",", …)

Parameters:

Notes

- file: the path to the file containing the data to be imported into R.
- header: logical value. If TRUE, read.csv2() assumes that your file has a header row, so row 1 is the name of each column. If that's not the case, you can add the argument header = FALSE.
- sep: the field separator character
- dec: the character used in the file for decimal points.

Example:

- R

```
# R program to read a file in table format
 # Using read.csv2()
myData = read.csv2("basic.csv")
print(myData)
```

Output:

```
Name.Age.Qualification.Address
1            Amiya,18,MCA,BBS
2             Niru,23,Msc,BLS
3             Debi,23,BCA,SBP
4             Biku,56,ISC,JJP
```

file.choose(): You can also use file.choose() with read.csv() just like before.

Example:

- R

```
# R program to read a file in table format
 # Using file.choose() inside read.csv()
myData = read.csv(file.choose())
# If you use the code above in RStudio
# you will be asked to choose a file
print(myData)
```

Output:

```
Name Age Qualification Address
1   Amiya  18      MCA     BBS
2   Niru  23         Msc     BLS
3   Debi  23         BCA     SBP
4   Biku  56         ISC     JJP
```

read_csv(): Additionally, this technique is employed to read comma (",") separated values with the aid of the readr package.

Syntax: read_csv(file, col_names = TRUE)

Notes

Parameters:

- file: the path to the file containing the data to be read into R.
- col_names: Either TRUE, FALSE, or a character vector specifying column names. If TRUE, the first row of the input will be used as the column names.

Example:

- R

```
# R program to read a file in table format
# using readr package
# Import the readr library
library(readr)
 # Using read_csv() method
myData = read_csv("basic.csv", col_names = TRUE)
print(myData)
```

Output:

```
Parsed with column specification:
cols(
  Name = col_character(),
  Age = col_double(),
  Qualification = col_character(),
  Address = col_character()
)
# A tibble: 4 x 4
  Name    Age Qualification Address
1 Amiya    18 MCA           BBS
2 Niru     23 Msc           BLS
3 Debi     23 BCA           SBP
4 Biku     56 ISC           JJP
```

Reading a file from the internet

It's possible to use the functions read.delim(), read.csv() and read.table() to import files from the web.

Example:

- R

```
# R program to read a file from the internet
 # Using read.delim()
myData = read.delim("http://www.sthda.com/upload/boxplot_format.txt")
print(head(myData))
```

Notes

Output:

Nom	variable	Group
1 IND1	10	A
2 IND2	7	A
3 IND3	20	A
4 IND4	14	A
5 IND5	14	A
6 IND6	12	A

Programming in R One of the most potent languages utilised specifically for data analytics in a variety of sectors is language. Reading and writing data from different files, such as Excel, CSV, text files, etc., is a requirement for data analysis.

R – Writing to Files

Writing Data to CSV files in R Programming Language

CSV stands for Comma Separated Values. There is a lot of statistical data handled by these files. Following is the syntax to write to a CSV file:

Syntax:

R

write.csv(my_data, file = “my_data.csv”)

write.csv2(my_data, file = “my_data.csv”)

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

- write.csv() uses “.” for the decimal point and a comma (“, ”) for the separator.

write.csv2() uses a comma (“, ”) for the decimal point and a semicolon (“;”) for the separator.

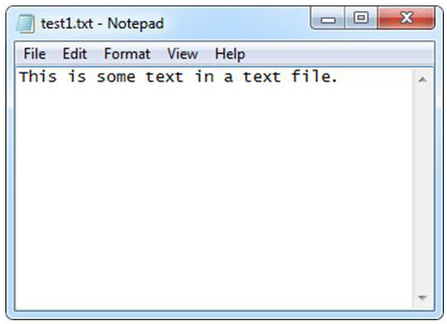
Writing Data to text files

As a first step towards a “Paperless World,” text files are frequently employed in practically all applications in our daily lives. Writing to.txt files is rather comparable to writing to CSV files. Following is the syntax to write to a text file:

Syntax:

R

write.table(my_data, file = “my_data.txt”, sep = “”)



Notes

Writing Data to Excel files

Installing the "xlsx package" is required to write data to Excel; this package essentially provides a Java-based means of reading, writing and committing changes to Excel files. It can be installed as follows:

install.packages("xlsx")

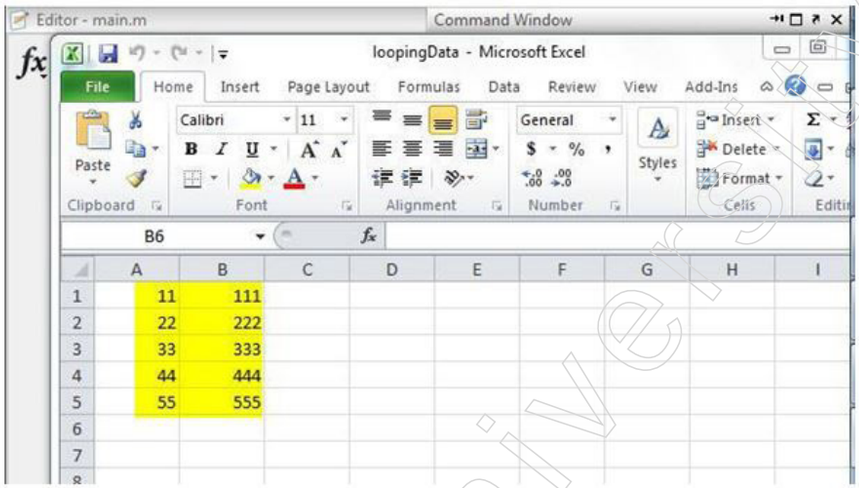
and can be loaded and General syntax of using it is:

R

library("xlsx")

write.xlsx(my_data, file = "result.xlsx",

sheetName = "my_data", append = FALSE).

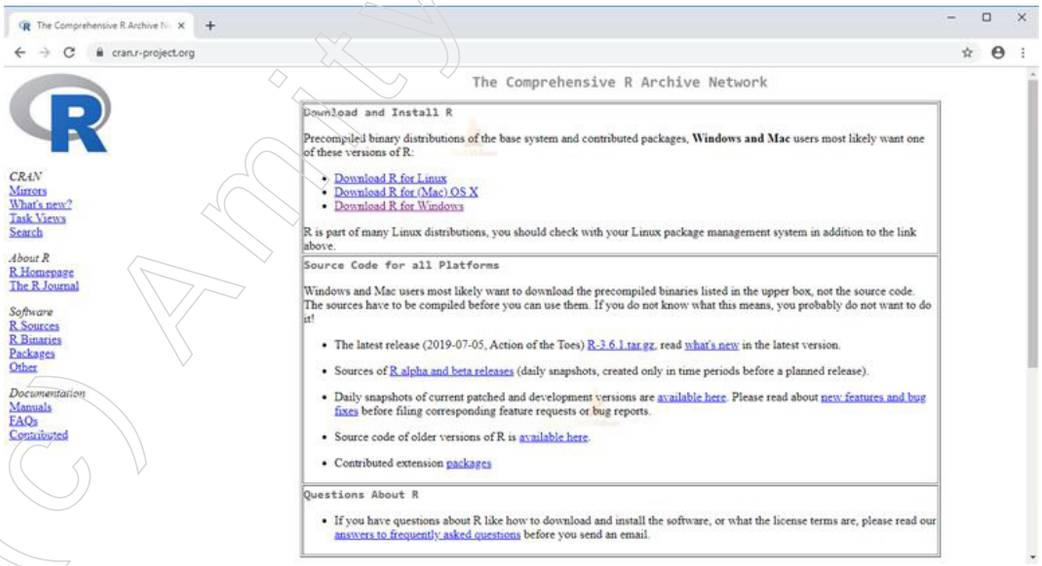


2.1.4 Installing "R" on Windows or Mac

To install R and RStudio on windows, go through the following steps:

Install R on windows

Step – 1: Go to CRAN R project website.

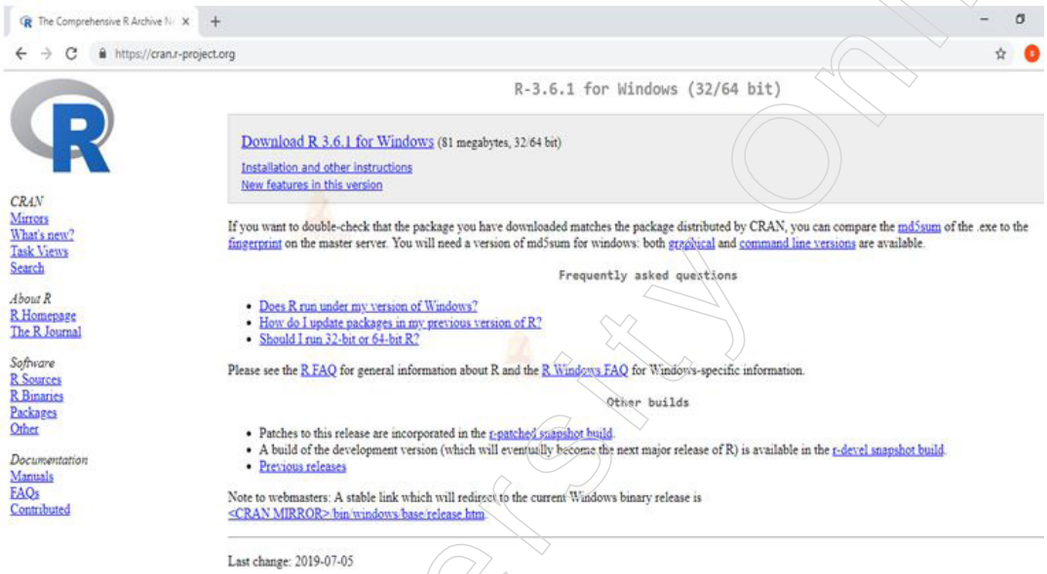


Notes

Step – 2: Click on the Download R for Windows link.

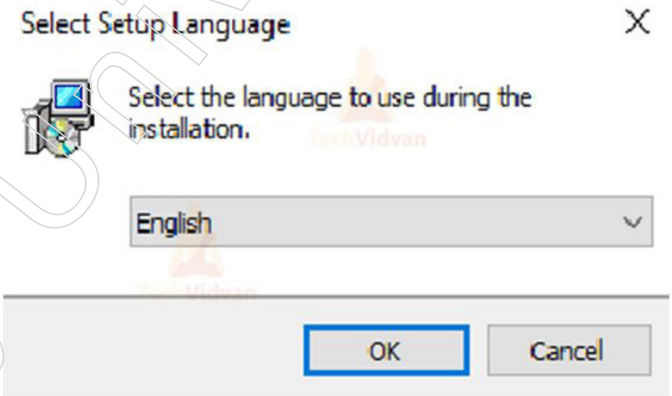
Step – 3: Click on the base subdirectory link or install R for the first-time link.

Step – 4: Click Download R X.X.X for Windows (X.X.X stand for the latest version of R. e.g.: 3.6.1) and save the executable .exe file.



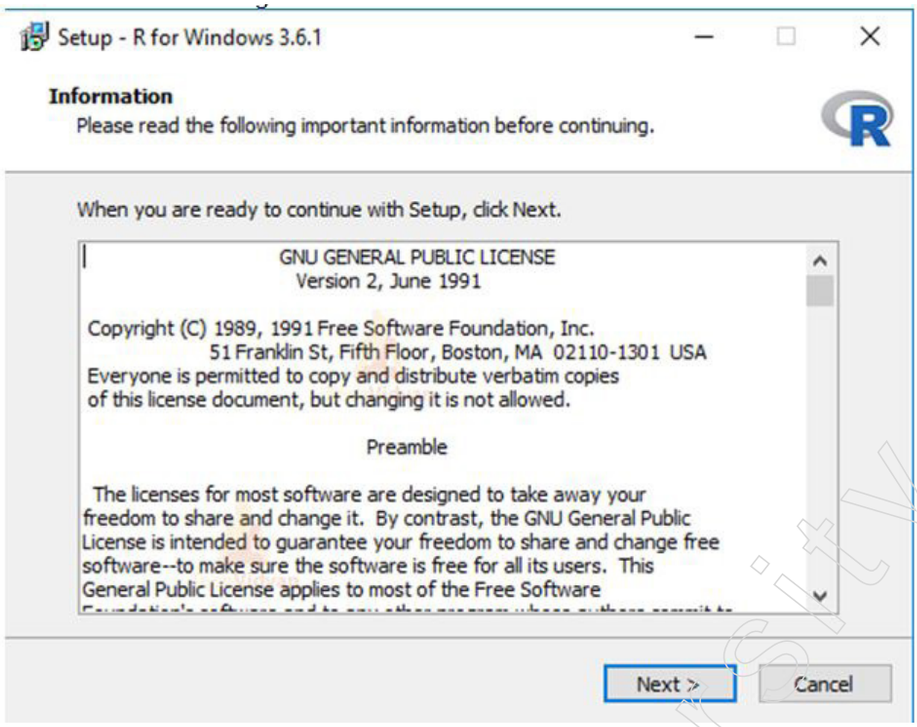
Step – 5: Run the .exe file and follow the installation instructions.

5.a. Select the desired language and then click Next.

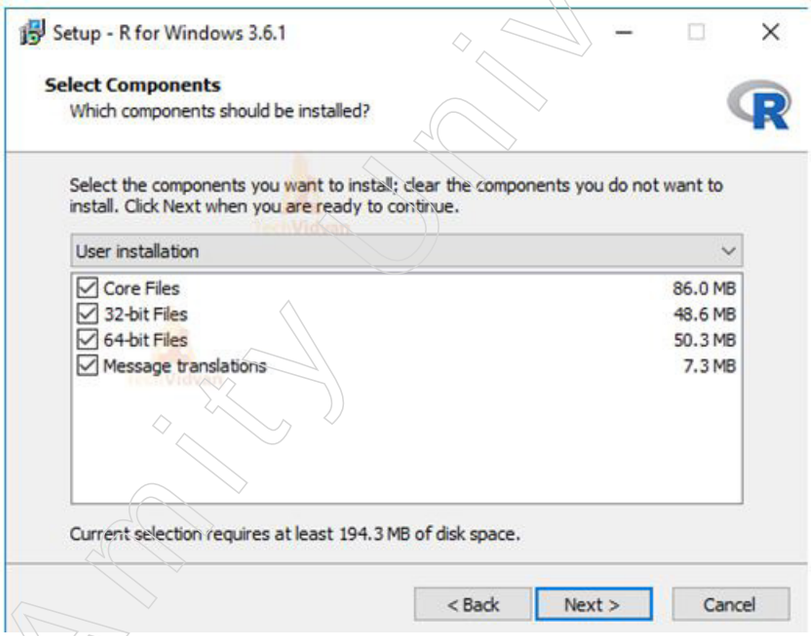


5.b. Read the license agreement and click Next.

Notes

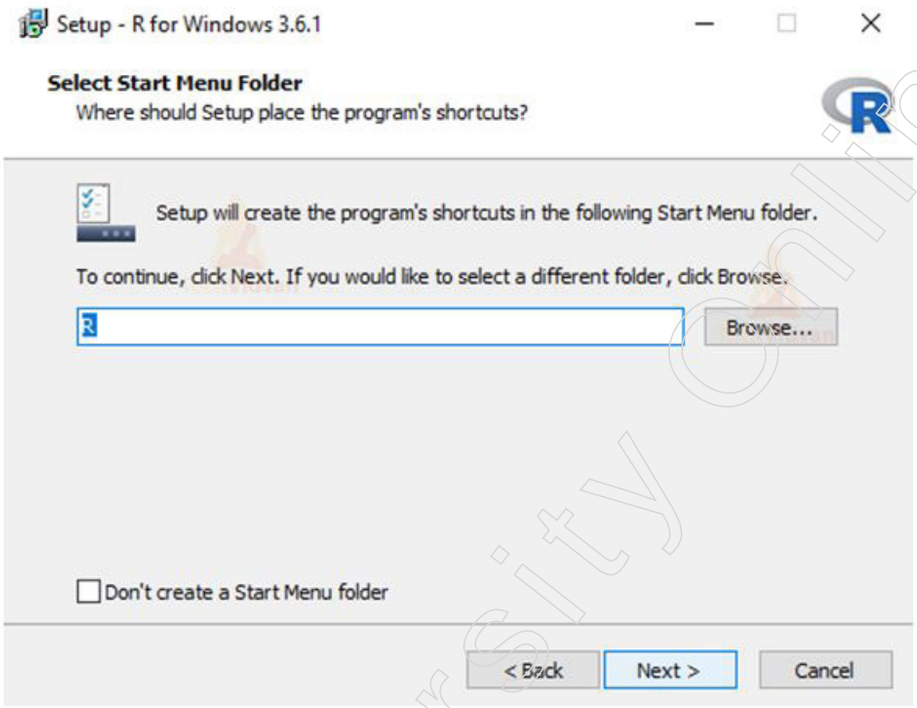


5.c. Select the components you wish to install (it is recommended to install all the components). Click Next.

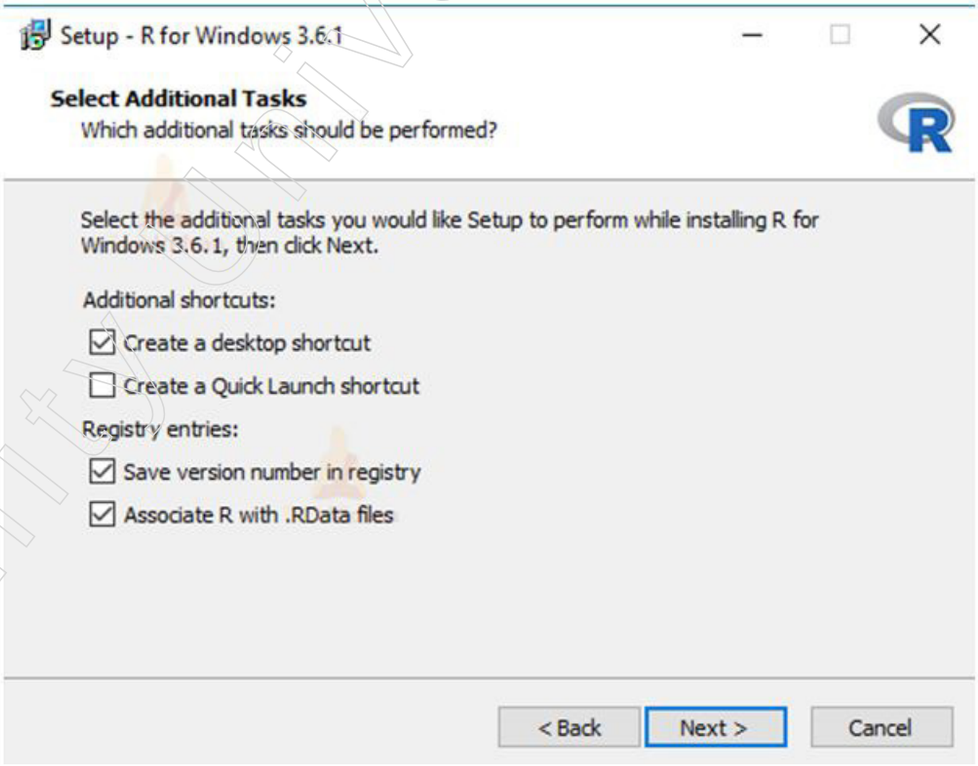


5.d. Enter/browse the folder/path you wish to install R into and then confirm by clicking Next.

Notes

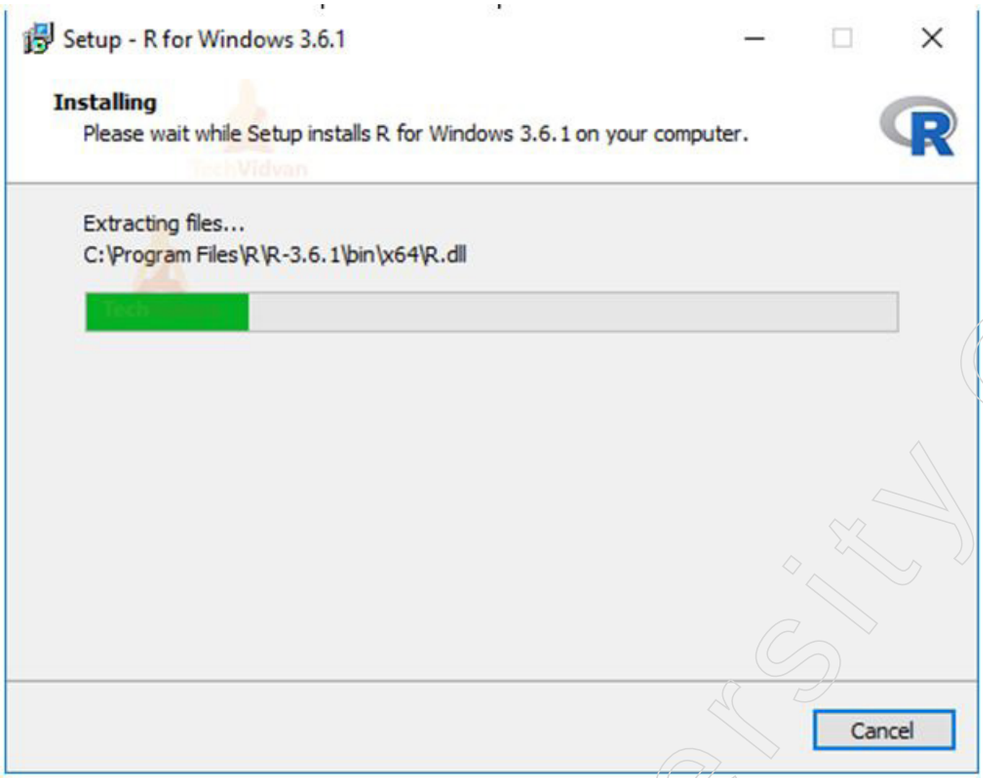


5.e. Select additional tasks like creating desktop shortcuts etc. then click Next.

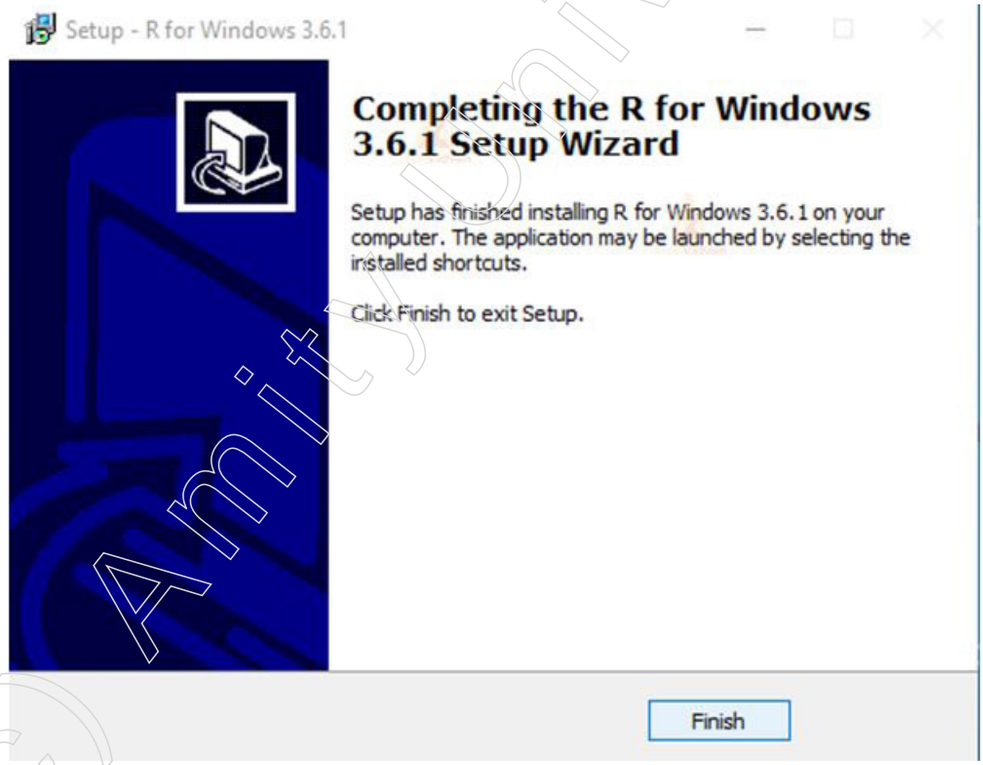


5.f. Wait for the installation process to complete.

Notes



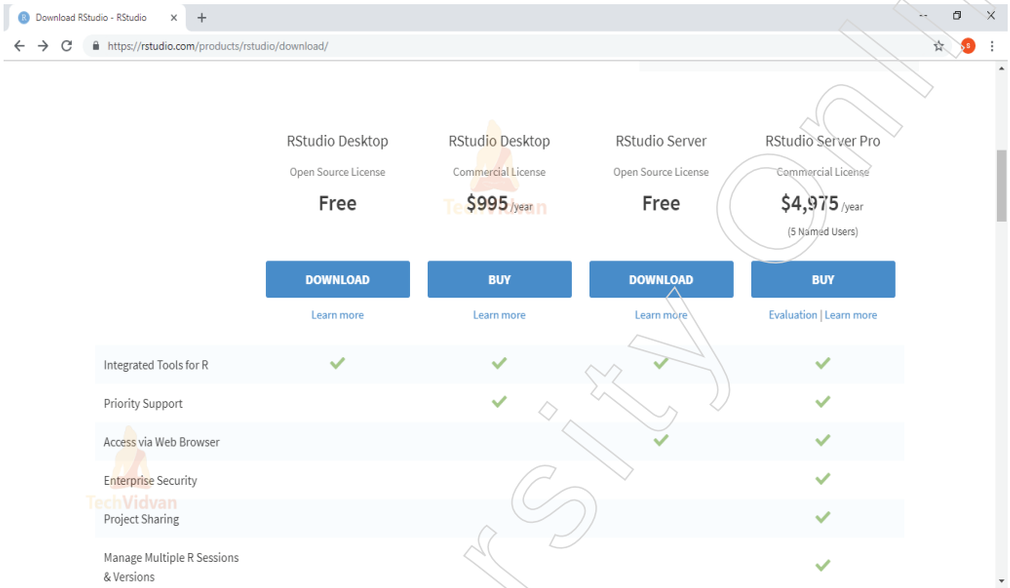
5.g. Click on Finish to complete the installation.



Notes

Install RStudio on Windows

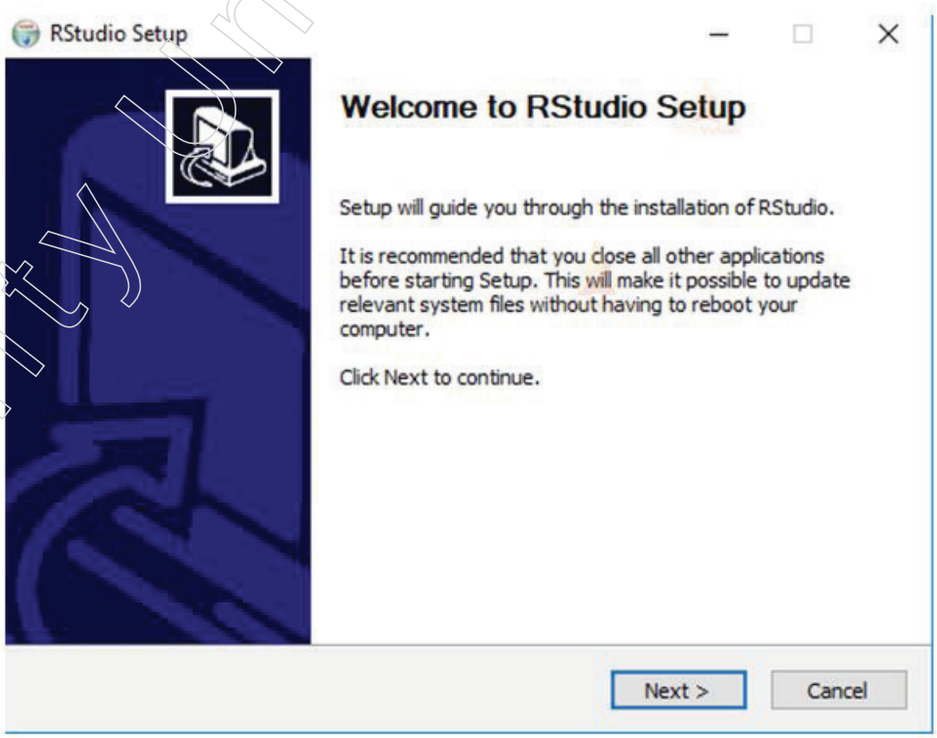
Step – 1: Let's install RStudio when R-base has been set up. Start by visiting the download RStudio page and selecting the RStudio desktop download option.



Step – 2: Save the.exe file by clicking the link for the RStudio for Windows version.

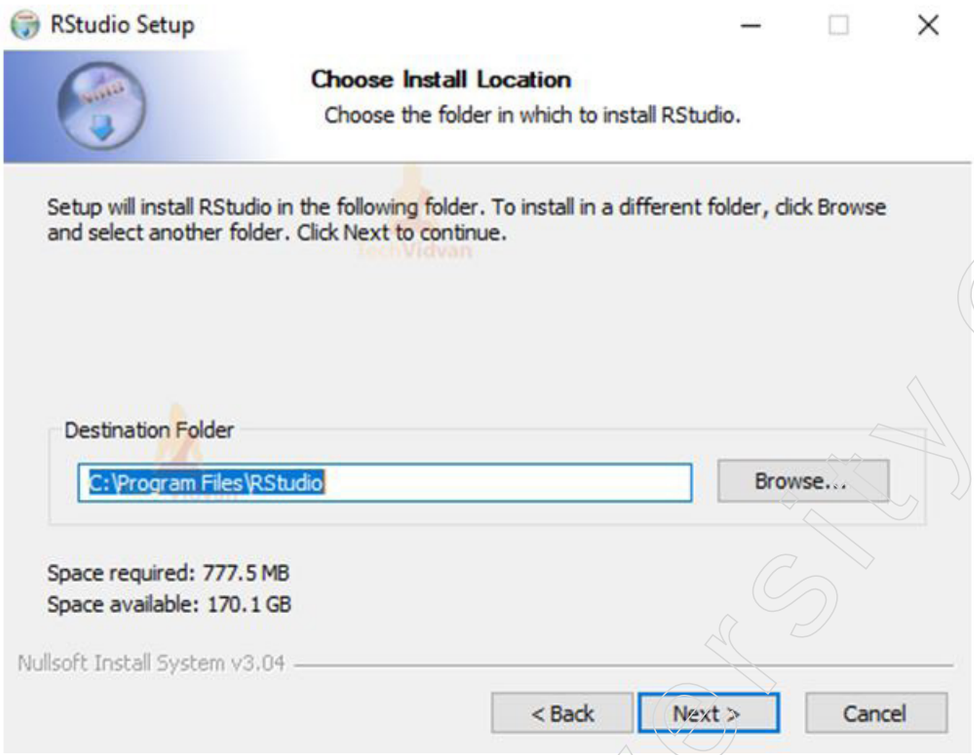
Step – 3: Run the.exe and adhere to the installation guidelines.

3.a. Click Next on the welcome window.

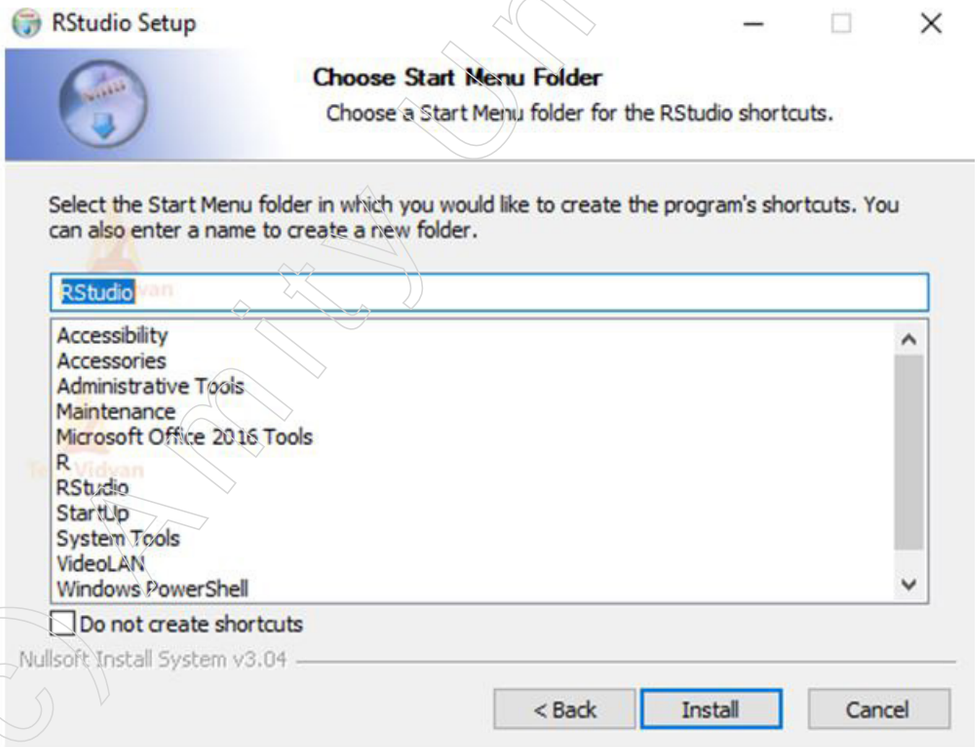


Notes

3.b. To continue, type or go to the installation folder, then click Next.

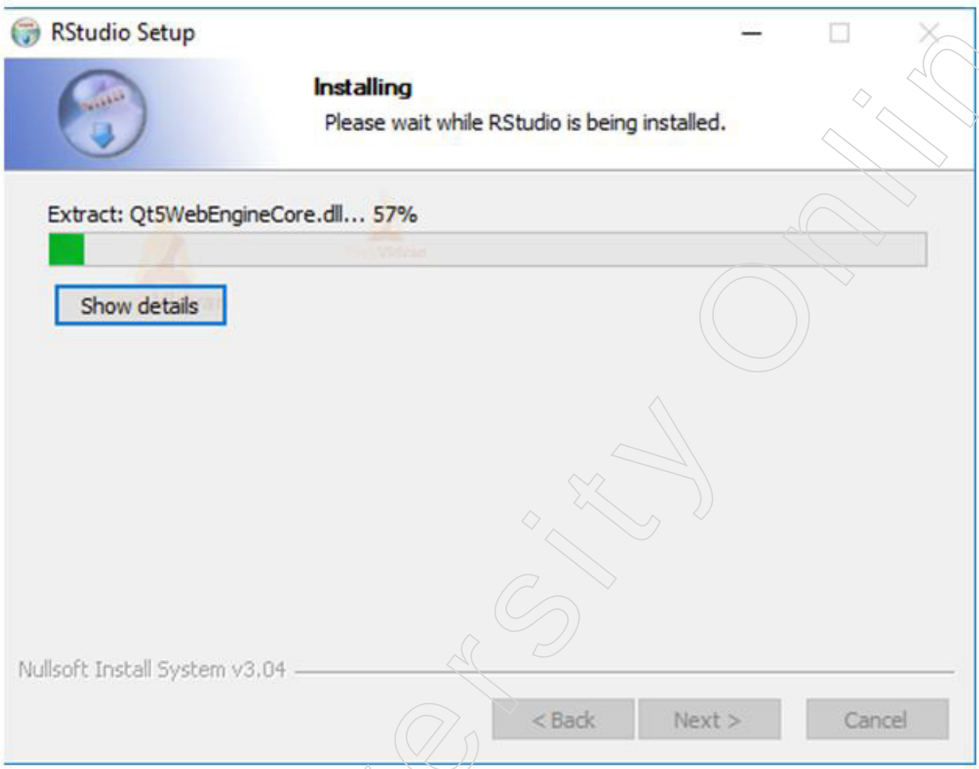


3.c. Click Next after choosing the folder for the start menu shortcut or choosing not to create shortcuts.

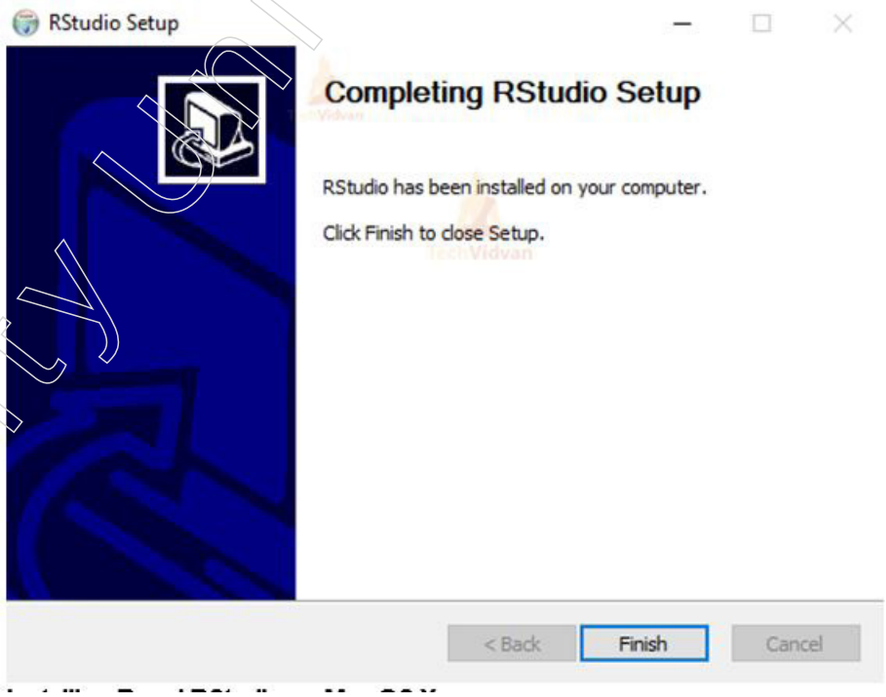


Notes

3.d. Hold off until the installation is finished.



3.e. To complete the installation, click Finish.



Installing R and RStudio on Mac OS X

Follow these instructions to install R and RStudio on Mac OS X:

Notes

Install R on Mac

Step – 1: Go to CRAN R Project Website.

Step – 2: Click on the Download for (Mac) OS X link.

Step – 3: Click on the link for the pkg file of the latest R version and save it.

Step – 4: Double click the downloaded file and follow installation instructions.

Install RStudio on Mac OS X

Step – 1: With the r-base installed, you need to install RStudio. To do that, go to download RStudio and click on the download button for the RStudio desktop.

Step – 2: Click on the link for the Mac OS X version of RStudio and save the .dmg file.

Step – 3: Drag and drop the downloaded file into your programmes folder after double clicking it.

Now that R and RStudio have been installed on your computer, let's look at a few packages that could aid in learning R and maximising its use.

Some useful Packages in R

There are packages available on CRAN for anything you would require when working with R and it is constantly expanding. These packages contain a lot of the useful R functions. Simply enter the following command in RStudio to install a package:

> install.packages("<package name>")

Once installed, a package can be made available in the current R session using the command:

> library("<package name>")

Here are a few packages that are well-known for their dependability and effectiveness, despite the fact that the sheer quantity of options offered might be perplexing at times:

- Tidyverse – A group of programmes called Tidyverse combine their efforts to clean, analyse, model and visualise data. Packages like ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr and forcats are included in the core package of tidyverse.
- Installr – With only one command, installr enables you to upgrade R and all of its packages.
- Rtweet – Twitter is the main platform for extracting tweets and creating sentiment analysis and sentiment prediction models. You can perform sentiment analysis and scrape Tweets using the rtweet programme.
- MLR (Machine Learning in R) – The MLR package enables you to carry out a variety of machine learning activities. All the well-known machine learning techniques applied in ML projects are included in MLR.
- Reticulate – Reticulate enables the usage of Python in the R environment alongside R. Additionally, you can utilise significant Python libraries within R itself.
- R markdown – R markdown allows you to write papers in a variety of file types, including pdf, HTML and MS Word documents, while containing R codes, outcomes and visualisations to make detailed reports.
- Shiny – An R package called Shiny enables you to create interactive web

Notes

applications. You can include the results of your investigation into the web-apps by using shine. This makes it possible for consumers to interact with your data and findings to gain a better understanding, which enhances the results' communication.

You can create interactive web applications using the R package Shiny. The results of your analysis can be integrated into the web-apps using shiny. Users can interact with your data and findings in this way to have a better understanding, which enhances the results' communication.

2.1.5 Data Types for R

In computer programming, R data types are utilised to define the sorts of data that can be stored in a variable. The proper data type must be chosen for efficient memory usage and accurate computation. Each type of R data has its own set of rules and limitations.

Each variable in R has an associated data type. Each R-Data Type may perform a different set of operations on it and has a different memory need. The following basic R-data types are available in R programming language and the accompanying table lists each data type's possible values.

Basic Data Types	Values	Examples
Numeric	Set of all real numbers	"numeric_value <- 3.14"
Integer	Set of all integers, Z	"integer_value <- 42L"
Logical	TRUE and FALSE	"logical_value <- TRUE"
Complex	Set of complex numbers	"complex_value <- 1 + 2i"
Character	"a", "b", "c", …, "@", "#", "$", …., "1", "2", …etc	"character_value <- "Hello Geeks"
Raw	as.raw()	"single_raw <- as.raw(255)"

Let's discuss each of these R data types one by one.

1. Logical Data Type

In R, the logical data type is also referred to as the Boolean data type. TRUE and FALSE are the only two possible values. For example,

```
bool1 <- TRUE

print(bool1)
print(class(bool1))

bool2 <- FALSE

print(bool2)
print(class(bool2))
```

Output

```
[1] TRUE
[1] "logical"
```

Notes

[1] FALSE

[1] “logical”

In the above example,

bool1 has the value TRUE,

bool2 has the value FALSE.

Note: You can also define logical variables with a single letter - T for TRUE or F for FALSE. For example,

```
is_weekend <- F
print(class(is_weekend))  # “logical”
```

2. Numeric Data Type

All real numbers in R, whether they have decimal values or not, are represented by the numeric data type. For example,

```
# floating point values
weight <- 63.5

print(weight)
print(class(weight))

# real numbers
height <- 182

print(height)
print(class(height))
```

Output

[1] 63.5

[1] “numeric”

[1] 182

[1] “numeric”

Here, both weight and height are variables of numeric type.

3. Integer Data Type

With no decimal points, actual values are specified by the integer data type. For integer data, specify it with the suffix L. For example,

```
integer_variable <- 186L
print(class(integer_variable))
```

Output

[1] “integer”

Notes

Here, 186L is an integer data. So, get “integer” when print the class of integer variable.

4. Complex Data Type

Purely imaginary values in R are specified using the complex data type. For the imaginary portion, indicate it using the suffix i. For example,

```
# 2i represents imaginary part
complex_value <- 3 + 2i

# print class of complex_value
print(class(complex_value))
```

Output

[1] “complex”

Here, 3 + 2i is of complex data type because it has an imaginary part 2i.

5. Character Data Type

A variable’s character data type is used to indicate character or string values.

A string is a collection of characters used in programming. For instance, “Apple” is a string while “A” is a single character.

You can use single quotes ‘’ or double quotes “” to represent strings. In general, use:

‘’ for character variables

“” for string variables

For example,

```
# create a string variable
fruit <- “Apple”

print(class(fruit))

# create a character variable
my_char <- ‘A’

print(class(my_char))
```

Output

[1] “character”

[1] “character”

Here, both the variables - fruit and my_char - are of character data type.

6. Raw Data Type

Values are specified as raw bytes in a raw data type. The following techniques can be used to change character data types into raw data types and vice versa:

charToRaw() - converts character data to raw data

Notes

rawToChar() - converts raw data to character data

For example,

```
# convert character to raw
raw_variable <- charToRaw("Welcome to Programiz")

print(raw_variable)
print(class(raw_variable))

# convert raw to character
char_variable <- rawToChar(raw_variable)

print(char_variable)
print(class(char_variable))
```

Output

```
[1] 57 65 6c 63 6f 6d 65 20 74 6f 20 50 72 6f 67 72 61 6d 69 7a
[1] "raw"
[1] "Welcome to Programiz"
[1] "character"
```

2.2 "R" Add-ins

What are RStudio add-ins? Extensions known as RStudio add-ins offer a quick method for running complex R functions from within RStudio. In other words, the corresponding code is executed without your involvement when you run an add-in (by clicking a button in the Add-ins menu). If you're still confused, keep in mind that when loading a dataset into RStudio, you have two options:

- importing it using coding (thanks to the read.csv() function for instance)
- or you can import it by clicking on the "Import Dataset" button in the Environment pane, set the importing settings, then click on "Import"

Like the Import Dataset button in RStudio, but with more common functionalities. Therefore, you could create code in the same way as you could import a dataset by writing code, but thanks to RStudio add-ins, you may execute code without writing the required code. You can have RStudio run the necessary code for you by utilising the add-ins.

A Shiny application that takes user input and uses it to make a plot is an example of a complicated RStudio add-in. RStudio add-ins can be as simple as a function that inserts a common piece of code. The benefit of using RStudio add-ins over writing intricate, advanced code yourself is that you can execute it considerably more quickly.

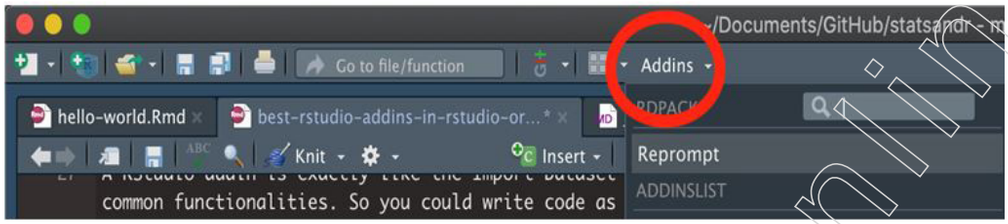
Installation

R packages are used to distribute RStudio Add-ins. Therefore, you must install them before using them.

The same method used to install a package can also be used to install an add-in: install.packages("name_of_addin"). The add-in will be immediately accessible

Notes

within RStudio via the Add-ins option at the top once you have installed the R package containing it.



2.2.1 Importing Data into R from Different File Formats

Data is an accumulation of facts. Data comes in a multiple format. A variety of file formats, including txt, CSV and other delimiter separated files, can be used to import data into R, before being analysed using the R programming language. After importing the data, modify, examine and report it.

Import CSV file into R

Method 1: Using read.csv() methods.

Here import csv file using read.csv() method in R.

Syntax: read.csv(path, header = TRUE, sep = ",")

Arguments:

path: The path of the file to be imported

header: By default: TRUE. Indicator of whether to import column headings.

sep = "," : The separator for the values in each row.

R

```
# specifying the path
path <- "/gfg.csv"

# reading contents of csv file
content <- read.csv(path)

# contents of the csv file
print (content)
```

Output:

```
ID  Name   Post    Age
1 5 H      CA      67
2 6 K      SDE     39
3 7 Z      Admin   28
```

Method 2: Using read.table() methods.

Here use read.table() methods to import CSV file into R Programming Language.

R

```
# simple R program to read csv file using read.table()
```

Notes

```
x <- read.csv2("D://Data//myfile.csv", header = TRUE, sep=", ")

# print x
print(x)
```

Output:

```
  Col1.Col2.Col3
1     100, a1, b1
2     200, a2, b2
3     300, a3, b3
```

Importing Data from a Text File

Using simple R functions, import or read.txt files with ease. read.table(). To read a file in table format, use the read.table() function. This feature is user-friendly and adaptable.

Syntax:

```
# read data stored in .txt file
x<-read.table("file_name.txt", header=TRUE/FALSE)
```

R

```
# Simple R program to read txt file
x<-read.table("D://Data//myfile.txt", header=FALSE)

# print x
print(x)
```

Output:

```
V1 V2 V3
1 100 a1 b1
2 200 a2 b2
3 300 a3 b3
```

If the header argument is set at TRUE, which reads the column names if they exist in the file.

Importing Data from a delimited file

R has a function read.delim() to read the delimited files into the list. The file is by default separated by a tab which is represented by sep="", that separated can be a comma(,), dollar symbol($), etc.

Syntax: read.delim("file_name.txt", sep="", header=TRUE)

R

```
x <- read.delim("D://Data//myfile.csv", sep="|", header=TRUE)

# print x
```

Notes

```
print(x)
# print type of x
typeof(x)
```

Output:

```
X.V1.V2.V3
1 1, 100, a1, b1
2 2, 200, a2, b2
3 3, 300, a3, b3

[1] “list
```

Importing Json file in R

Here, we’ll demonstrate how to import a JSON file into the R programming language using the rjson package.

R

```
# Read a JSON file

# Load the package required to read JSON files.
library(“rjson”)

# Give the input file name to the function.
res <- fromJSON(file = “E:\\exp.json”)

# Print the result.
print(res)
```

Output:

```
$ID
[1] “1” “2” “3” “4” “5”

$Name
[1] “Mithuna”  “Tanushree” “Parnasha”  “Arjun”  “Pankaj”

$Salary
[1] “722.5”  “815.2”  “1611”  “2829”  “843.25”
```

Importing XML file in R

Utilising the R programming language’s XML Package, import an XML file here.

XML file for demonstration:

HTML

```
<RECORDS>
<STUDENT>
```

<ID>1</ID>
<NAME>Alia</NAME>
<MARKS>620</MARKS>
<BRANCH>IT</BRANCH>
</STUDENT>
<STUDENT>
<ID>2</ID>
<NAME>Brijesh</NAME>
<MARKS>440</MARKS>
<BRANCH>Commerce</BRANCH>
</STUDENT>
<STUDENT>
<ID>3</ID>
<NAME>Yash</NAME>
<MARKS>600</MARKS>
<BRANCH>Humanities</BRANCH>
</STUDENT>
<STUDENT>
<ID>4</ID>
<NAME>Mallika</NAME>
<MARKS>660</MARKS>
<BRANCH>IT</BRANCH>
</STUDENT>
<STUDENT>
<ID>5</ID>
<NAME>Zayn</NAME>
<MARKS>560</MARKS>
<BRANCH>IT</BRANCH>
</STUDENT>
</RECORDS>

Reading XML file:

After installing the package and using the xmlparse() function to parse it, it can be read.

R

```
# loading the library and other important packages
library("XML")
```

Notes

```
library("methods")

# the contents of sample.xml are parsed
data <- xmlParse(file = "sample.xml")

print(data)
```

Output:

```
1
Alia
620
IT
2
Brijesh
440
Commerce
3
Yash
600
Humanities
4
Mallika
660
IT
5
Zayn
560
IT
```

Importing SPSS sav File into R

Here, we'll read an SPSS.sav file using the R programming language. We'll make use of the haven package for this. R's read_sav() function, found in the haven package, is used to read SPSS files.

Syntax: read_sav("FileName.sav")

```
R
# import haven library package
library("haven")

# Use read_sav() function to read SPSS file
dataframe <- read_sav("SPSS.sav")
```

dataframe

Output:

```
> # import haven library package
> library("haven")
> # Use read_sav() function to read SPSS file
> dataframe <- read_sav("SPSS.sav")
> dataframe
# A tibble: 5 x 3
  Batch Students Class
  <dbl>    <dbl> <chr>
1  2017     2300 DSA Essential
2  2018     1200 Placement100
3  2019     3500 C++: Expert
4  2020     1400 Web Development Bootcamp
5  2021      120 Android Development Bootcamp
>
```

Notes

2.2.2 Scrape Data from the Web

Web scraping is a method for turning unstructured data from HTML tags that are present on the web to a structured format that can be readily accessed and used. Many popular languages have methods for executing web scraping. Use web scraping to extract information from IMDB to provide ourselves practical experience.

Data extraction from the internet can be done in several ways. Among the common methods are:

- Human Copy-Paste: This method of web data scraping is both slow and effective. Individual people must analyse and copy the data to local storage as part of this.
- Text pattern matching: Using the regular expression matching capabilities of programming languages is another straightforward yet effective method for extracting data from the web.
- API Interface: Numerous websites, like Facebook, Twitter, LinkedIn and others, offer public and/or private APIs that can be used by calling them with the usual code to retrieve data in the required format.
- DOM Parsing: Programmes can retrieve dynamic content produced by client-side scripts utilising web browsers. Additionally, web pages can be parsed into a DOM tree, from which programmes can get portions of these pages.

2.2.3 Tidy Data Using Tidy verse

It is frequently stated that cleaning and preparing the data takes about 80% of the time in data analysis. Furthermore, it must be repeated numerous times throughout the course of the research if new issues are discovered or fresh data is gathered. It is not only a first step. This section focuses on a tiny but crucial part of data cleaning that one refers to as data tidying: organising datasets to support analysis.

The tidy data principles offer a consistent approach to arrange data values within a dataset. Initial data cleaning is made simpler by standards because you don't have to start from scratch each time. The tidy data standard has been created to make it easier to explore and analyse the data at first, as well as to make it simpler to create data analysis tools that are compatible with one another.

Notes

"Tidy datasets are all alike, but every messy dataset is messy in its own way" - Hadley Wickham

A tidy dataset is intended to be presented in a way that allows for orderly additional processing:

1. Each variable in the data set is placed in its own column
2. Each observation is placed in its own row
3. Each value is placed in its own cell

If you frequently use Excel (and Excel pivot tables), you might think of clean data as being very "pivot-friendly" data. Consider a situation where you wanted a pivot table, but the raw data comprised columns and rows of dimensions (such as Campaign in rows and Device Category in columns, with Sessions in the cells).

Although tidy data isn't always the simplest to read by humans, it is suitable for sending into other R functions, especially those that are part of the tidyverse.

Two bits of good news about tidy data:

- Data for both Google Analytics and Adobe Analytics typically exits the API in a neat manner.
- The tidyverse (the tidyr package) includes functions designed for transforming data to and from a tidy format.

The Tidyverse

The Hadleyverse, a collection of R packages developed or maintained by Hadley Wickham, has a new name: Tidyverse.

You can make tidy data with the aid of tidyr. Clean data are those that:

1. Every column is variable.
2. Every row is an observation.
3. Every cell is a single value.

The term "tidy data" refers to a common method of data storage that is applied as much as feasible throughout the tidyverse. You'll spend less time wrangling with the tools and more time focusing on your research if you make sure your data is organised.

These packages each assist the other in terms of concepts and output to cover the full range of data analysis inside R. Within the tidyverse, the Big Three packages are:

- dplyr - data manipulation of data frames
- tidyr - tools to tidy (and untidy) data frames
- ggplot2 - plotting tidy data

In addition, some other packages in the tidyverse that you may come across are:

- tibble - creating more user-friendly data.frames
- purrr - more general data manipulation tools
- magrittr - the origin of %>%, extensivly used in the tidyverse (This isn't actually a package developed by Wickham, but his packages have a dependency on this one, dropping, "Oh, yeah. The pipe in dplyr comes from the magrittr package" in casual conversation with R folk will provide a nominal degree of street cred.)

Notes

- broom - turn statistical models into tidy data frames/tiddles

The tidyverse has become so popular that you can now just install/load the 'tidyverse' package, which will install all these items (and a few more) all at once. This list only represents a small portion of the tidyverse packages.

2.2.4 Process Strings with Regular Expressions

Regular Expressions (also known as regex) are a collection of commands that match patterns and are used to find string sequences in enormous amounts of text. These instructions are flexible enough to handle any text or string class because they are made to match a family of text (alphanumeric, numbers, words).

In other words, regular expressions allow you to write shorter scripts while still extracting more value from text data.

A set of functions called "string manipulation" are used to extract data from text variables. These routines are frequently used in machine learning to do feature engineering, or the process of constructing new features from pre-existing string features. There are numerous string manipulation functions available in R packages like stringr and stringi.

R also includes several base functions for manipulating strings. These are designed to work in conjunction with regular expressions. The following are the practical distinctions between regular expressions and string manipulation functions:

1. To do straightforward actions like splitting a string or removing the first three letters, among other things, employ string manipulation routines. Regular expressions are used for more difficult jobs like extracting email IDs or dates from a collection of text.
2. The responses from string manipulation functions are predetermined. They stay true to their usual behaviour. whereas regular expressions can be altered however you like.

Any value that is contained in quotations ("") is a string in R. Yes, you can have strings that are numbers. R alerts strings that fall under the character class

Let's look at some of the commonly used base R functions (also available in stringr) to modify strings:

Functions	Description
nchar()	It determines how many characters are there in a string or vector. Str_length() is a suitable replacement function in the stringr library.
tolower()	A string is changed to lower case. You can also use the str_to_lower() function as an alternative.
toupper()	It changes the case of a string to upper case. You can also use the str_to_upper() function as an alternative.
chartr()	It is used to replace each character in a string. Alternatively, you can use str_replace() function to replace a complete string
substr()	A string's characters can be replaced one by one using it. As an alternative, you can replace a whole string using the str_replace() function.
setdiff()	It is used to establish how two vectors differ from one another.
setequal()	It is utilised to determine whether the string values in the two vectors match.
abbreviate()	It is used to shorten strings. The short string's length must be supplied.

Notes

strsplit()	It is utilised to divide a string depending on a standard. It gives back a list. You can also employ the str_split() method. You can use this function to turn your list output into a character matrix.
sub()	The first match in a string can be located and replaced using it.
gsub()	It is utilised to locate and swap out each match within a string or vector. You can also use the str_replace() method.

In addition to the function mentioned above, there are several other functions that were created specifically to deal with regular expressions (also known as regex). Yes, R is just as capable of processing text data. There are various ways to carry out a specific task in regex. You must therefore adhere to a specific approach when learning to prevent confusion.

The available base regex functions for employing regular expressions are grep(), grepl(), regexpr(), gregexpr(), regexec() and regmatches(). Here is a short summary of these features:

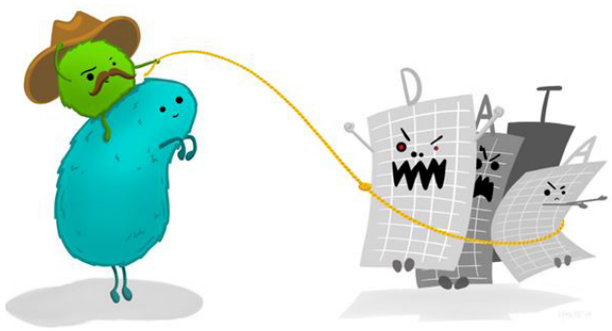
Functions	Description
Grep	returns the matching string's index or value
Grepl	returns the matching string's Boolean value (True or False)
Regexpr	give the first match's index back
Gregexpr	the index of all matches is returned
Regexec	is a hybrid of regexpr and gregexpr
Regmatches	the string that was matched at the specified index is returned. It is used in conjunction with regexpr and gregexpr

Regular expressions in R can be divided into 5 categories:

1. Metacharacters
2. Sequences
3. Quantifiers
4. Character Classes
5. POSIX character classes

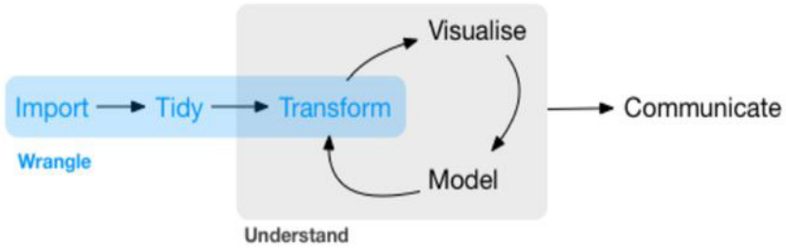
2.2.5 Wrangle Data Using Dplyr

The element of any data analysis job that requires the most time is typically data wrangling. Although it might not always be enjoyable, it is the basis for all the labour that comes after. Additionally, data wrangling frequently takes a lot longer than data analysis or data visualisation.



Notes

Import, cleaning and transformation of data are all steps in the data wrangling process. The method iteratively contributes directly to the understanding or modelling side of data exploration. In a broader sense, data wrangling refers to the manipulation or integration of datasets for analytical purposes. This process frequently requires repetition as your understanding of the data and your demands for modelling and visualisation evolve.



No one wrangles for the sake of wrangling (usually), so the process always begins by answering the following two questions:

1. What do the input data looks like?
2. What should the output data look like given what one wants to do?

Common operations that occur during data wrangling:

- ❖ select specific variables
- ❖ filter observations by some criteria
- ❖ Add or modify (mutate) existing variables
- ❖ rename variables
- ❖ arrange rows by a variable
- ❖ summarise a variable conditional on others

A fundamental package from the "tidyverse" collection of packages, "dplyr" offers simple tools for these typical data manipulation tasks. Dplyr's guiding principle is that each function only does one specific task, as indicated by its name.

2.3 R- Controls and Functions

According to the conditions specified in the statements, control statements are expressions that are used to control the execution and flow of the programme. Using these structures, decisions are made after evaluating the variable.

When you wish to carry out a specific task repeatedly, functions are helpful. A function executes legitimate R commands that are contained within the function to create the output after accepting input arguments. The function name and the file in which the function is created need not match when using the R programming language and you can have one or more functions in R.

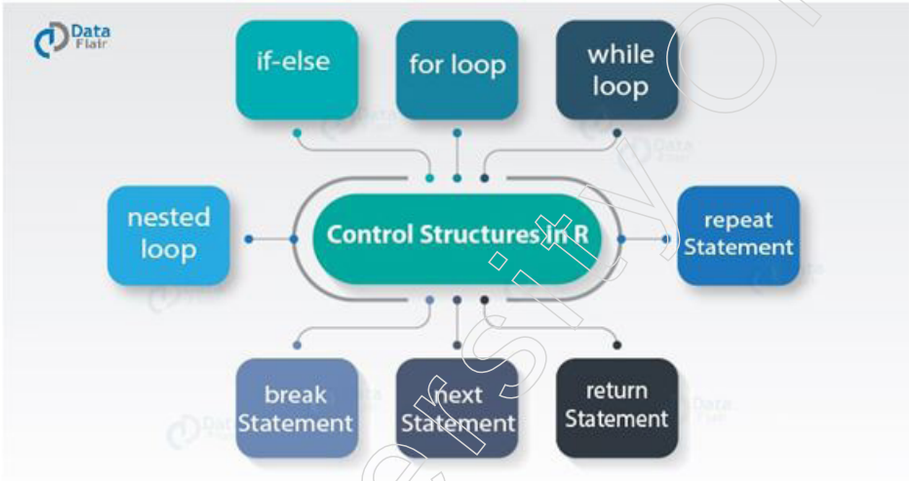
2.3.1 Control Structures

The control structures in R to regulate how expressions are executed. In R, these regulatory elements are also known as loops. R has eight different types of control structures:

- ❖ if

Notes

- if-else
- for
- nested loops
- while
- repeat and break
- next
- return



1. if Condition in R

Only if this condition is returned as TRUE will this task be completed. It is even simpler with R: You can omit the term "then" and indicate your preference in an if statement instead.

Syntax:

```
if (test_expression) {
statement
}
```

Example:

```
values <- 1:10
if (sample(values,1) <= 10)
print(paste(values, "is less than or equal to 10"))
```

2. if-else Condition in R

The parts of a "if...else" statement are like those in a "if" statement plus a few more:

- The keyword else, placed after the first code block.
- The second block of code, contained within braces, that has to be carried out, only if the result of the condition in the if() statement is FALSE.

Syntax:

```
if (test_expression) {
statement
```

Notes

```
} else {
statement
}
```

Example:

```
val1 = 10 #Creating our first variable val1
val2 = 5 #Creating second variable val2
if (val1 > val2){ #Executing Conditional Statement based on the comparison
print("Value 1 is greater than Value 2")
} else if (val1 < val2){
print("Value 1 is less than Value 2")
}
```

3. for Loop in R

A loop is a set of instructions that are repeated repeatedly until a particular condition is met. Loops are created by combining the verbs for, while and repeat with the additional clauses break and next.

Example:

These control structures in R, made of the rectangular box 'init' and the diamond. It is executed a known number of times. for is a block that is contained within curly braces.

```
values <- c(1,2,3,4,5)
for(id in 1:5){
print(values[id])
}
```

4. Nested Loop in R

It is like the conventional for loop, making the transition to a foreach loop simple. Foreach does not demand that the body of the for loop be converted into a function, in contrast to many parallel programming libraries for R. Because it is used to construct nested foreach loops, refer to this as a nesting operator.

Example:

```
mat <- matrix(1:10, 2)
for (id1 in seq(nrow(mat))) {
for (id2 in seq(ncol(mat))) {
print(mat[id1, id2])
}
}
```

5. while Loop in R

While(cond) expr is the format, where cond is the test condition and expr are an expression.

Notes

R would complain about the absence of the expression that should have provided the necessary True or False, but the truth is that it did not understand the word "response" before employing it in the loop. Additionally, accomplish this because the loop will not be run at all if correctly respond on the first try.

Example:

```
val = 2.987
while(val <= 4.987) {
val = val + 0.987
print(c(val,val-2,val-1))
}
```

6. repeat and break Statement in R

To halt iterations and move control outside of a loop (repeat, for, while), utilise a break statement inside the loop. This statement exits the deepest loop that is being evaluated in a nested looping scenario when there is a loop inside another loop.

A repeat loop is utilised to repeatedly iterate over a block of code. To break a repeat loop, there is no condition check. Ourselves expressly include a condition inside the loop's body and use the break statement to end it. Failure to do so will lead to an endless loop.

Syntax:

```
repeat {
# simulations; generate some value have an expectation if within some range,
# then exit the loop
if ((value - expectation) <= threshold) {
break
}
}
```

The repeat loop is an infinite loop and is used in association with a break statement.

Example:

Below, the code shows a repeat statement in R:

In a loop, a break statement is used to halt iterations and move control outside of the loop.

Example of Repeat Statement in R:

```
val <- 5
repeat {
print(val)
val <- val+1
if (val == 10){
break
```

```
}
}
```

Example of Break Statement in R:

```
values = 1:10
for (id in values){
if (id == 2){
break
}
print(id)
}
```

7. next Statement in R

Using the next statement, a cycle is skipped and the next one is entered instead. To analyse the restriction that keeps the current loop in place, it goes ahead to that section. You can skip the current loop iteration with the following statement without breaking the loop.

Example:

```
x = 1: 4
for (i in x) {
if (i == 2) {
next
}
print(i)
}
```

8. return Statement in R

Frequently need certain functions to carry out processing and return the results. This is achieved using R's return() command.

Syntax:

```
return(expression)
```

Example:

```
check <- function(x) {
if (x > 0) {
result <- "Positive"
} else if (x < 0) {
result <- "Negative"
} else {
result <- "Zero"
```

Notes

```
}
return(result)
}
```

In the console window, type:

```
> check(1)
> check(-10)
> check(0)
```

2.3.2 R Functions

Simply put, R function is a chunk of code that may be called and run from any other place in your programme. They are utilised to break down our code into digestible bits and avoid using repeating codes. You can pass data into functions using arguments and have them return a different set of data. Use the function() reserve keyword in R to define a function. The grammar is:

```
func_name <- function (parameters) {
statement
}
```

Here, func_name is the name of the function. For example,

```
# define a function to compute power
power <- function(a, b) {
 print(paste("a raised to the power b is: ", a^b))
}
```

Here, we've defined the power function, which has two inputs: a and b. have code to output the value of a raised to the power b inside the function.

Call the Function

Once the function has been defined, you can use the function name and arguments to call it. For example,

```
# define a function to compute power
power <- function(a, b) {
 print(paste("a raised to the power b is: ", a^b))
}

# call the power function with arguments
power(2, 3)
```

Here, the function has been called with the inputs 2 and 3. This will output 8 as the result of 2 increased to the power of 3.

Formal arguments are those that are used in the actual function. They may also be

known as parameters. Actual arguments are the values that are supplied to the function when it is called.

Notes

Named Arguments

The arguments given in the power() function call above must be passed in the same order as the parameters passed in the function definition.

This means that when call power(2, 3), the values a and b are given the values 2 and 3, respectively. You can use named arguments to alter the order in which parameters are supplied. For example,

```
# define a function to compute power
power <- function(a, b) {
print(paste("a raised to the power b is: ", a^b))
}

# call the power function with arguments
power(b=3, a=2)
```

No matter what order you send the arguments in during the function call, the outcome is the same in this case.

You can also use a mix of named and unnamed arguments. For example,

```
# define a function to compute power
power <- function(a, b) {
print(paste("a raised to the power b is: ", a^b))
}

# call the power function with arguments
power(b=3, 2)
```

Default Parameters Values

Functions can be given default parameter values. To do this, you can during function definition indicate an acceptable value for the function parameters.

When a function is called without an argument, the default setting is applied. For example,

```
# define a function to compute power
power <- function(a = 2, b) {
print(paste("a raised to the power b is: ", a^b))
}

# call the power function with arguments
power(2, 3)
# call function with default arguments
```

Notes

```
power(b=3)
```

Here, have just given the b argument as a named argument in the second call to the power() function. In this instance, the function definition's default value for an is used.

Return Values

The return() keyword can be used to retrieve values from a function. For example,

```
# define a function to compute power
power <- function(a, b) {
 return (a^b)
}

# call the power function with arguments
print(paste("a raised to the power b is: ", power(2, 3)))
```

Here, returned a^b rather than printing the result inside the function. The result is returned when the power() function is called with arguments and it can be printed right away.

Nested Function

There are two ways to build nested functions in R.

- By using one function inside of another.
- By enclosing one function inside another.

2.3.3 Work with Dates and Times as File Formats

Numerous functions in the R programming language deal with date and time. The date can be formatted and converted from one form to another using these procedures. R offers several format specifiers, which are included in the table below-

Specifier	Description
%a	Abbreviated weekday
%A	Full weekday
%b	Abbreviated month
%B	Full month
%C	Century
%y	Year without century
%Y	Year with century
%d	Day of month (01-31)
%j	Day in Year (001-366)
%m	Month of year (01-12)
%D	Data in %m/%d/%y format
%u	Weekday (01-07) Starts on Monday

Note: R offers a function named sys.Date() that returns the current date in order to obtain the Today date.

Notes

2.3.4 Scoping Rules

How R utilises the search list to associate a value with a symbol is connected to the scoping rules.

Consider the following function:

```
f <- function(x, y) {
 x^2 + y / z
}
```

The formal arguments for this function are x and y. Another symbol, z, appears in the function's body. Here, z is referred to as a free variable.

The way values are assigned to free variables is governed by the scoping rules of a language. Free variables are neither local variables (assigned within the function body) nor formal arguments.

Lexical scoping, or static scoping, is used in R. Dynamic scoping is an alternative to lexical scoping that some languages use. Lexical scoping proves to be especially beneficial for streamlining statistical calculations.

Lexical scoping in R means that:

The environment where the function was defined is searched for free variable values.

A closure or function closure is made up of a function and its enclosing environment (the context in which the function was defined, also known as enclosure).

It is possible to develop functions that "carry around" data using the function closure concept.

2.4 Loop Functions

To make things easier, R includes some incredibly helpful functions that implement looping in a condensed manner. Inherently vectorized functions make up the extremely large and robust family of apply functions. These R functions let you apply a certain function to a group of objects, such as files, vectors, matrices, or dataframes. They include:

1. lapply(): Loop over a list and evaluate a function on each element
2. sapply(): Same as lapply but try to simplify the result
3. apply(): Apply a function over the margins of an array
4. tapply(): Apply a function over subsets of a vector
5. mapply(): Multivariate version of lapply

There is another function called split() which is also useful, particularly in conjunction with lapply.

2.4.1 lapply Function

The lapply() function does the following simple series of operations:

1. It loops around a list, going through each item in turn.
2. It applies a function (that you provide) to each entry of the list.

Notes

3. and returns a list.

To get the help file type the following code.

?lapply()

Simply entering lapply into your console will display the lapply() function's body here.

lapply

2.4.2 sapply Function

The sapply() function returns an array or matrix object with the same length when used on a list, vector, or data frame. The R language's sapply() function converts a list, vector, or data frame into an object that is an array or matrix. Since the sapply() function applies a specified operation to every component of the object, it is not necessary to specify a MARGIN. The type of return object is the only distinction between it and lapply().

sapply(x, fun)

2.4.3 mapply Function

The mapply() function is a kind of multivariate apply that simultaneously applies a function to a number of inputs. If you wish to iterate over many R objects simultaneously, mapply() is the function for you. lapply() iterates over a single R object.

There is a way to vectorize the call of a non-vectorized function thanks to mapply. It is sapply in a multivariate form. mapply applies FUN to the first, second, third and so forth parts of each... argument. When required, arguments are repeated.

Get the help file by typing ?mapply in your R console. To get the list of arguments it takes just type str(mapply).

?mapply

str(mapply)

2.4.4 tapply Function

Apply a function to each (non-empty) cell of a ragged array, or to each group of values produced by a certain arrangement of the levels of a few different components. In essence, tapply() performs an action or function on a subset of the vector that has been divided by the specified factor variable.

To get the help file type the code.

?tapply

To see the arguments of tapply() function type str(tapply) in the console.

str(tapply)

2.4.5 Split Function

Using a factor or list of factors, the split() function divides a vector or other object into groups. The fundamental concept is that you can apply a function on subsets of a data structure that have been divided into groups according to another variable.

You can get the help file by typing

Notes

?spilt

The arguments of split() can be shown by just typing split in your R console.

split

2.5 Other "R" tools

The most potent and popular programming language for computational statistics, visualisation and data science is R, which is utilised in analytics tools. R is a popular data analysis tool that many statisticians and data scientists use to address issues in fields ranging from computational biology to quantitative marketing. The GNU project R is more comparable to the S language. It is regarded as a S language dialect.

R offers a variety of statistical and graphical approaches for data analysis. R is quite flexible and gives researchers a wide range of options. Runs on a variety of operating systems, including MacOS, Windows, UNIX and Linux, R is free software. It is also the most successful and widely used language and it has certain strong features. Companies with a high reputation for quality utilise it, including Google, Facebook, Shell, Merck, Bank of America, Pfizer and LinkedIn.

2.5.1 Basic Debugging Tools

Debugging is the process of clearing a program's code of errors so that it can function properly. When writing codes, some errors or issues inevitably surface during compilation and are more challenging to address. Therefore, correcting it requires a lot of time and several calls at different levels.

R uses warnings, messages and errors to help in debugging. In R, function debugging is considered debugging. Several debugging features are:

1. Editor breakpoint
2. traceback()
3. browser()
4. recover()

1. Editor Breakpoints

Editor By clicking to the left of the line in RStudio or pressing Shift+F9 while the cursor is on your line, breakpoints can be inserted. Similar to browser(), a breakpoint doesn't need modifying any code. Breakpoints are identified by a red circle on the left, indicating that if the source is executed, debug mode will be activated at this line.

2. traceback() Function

To provide detailed information on how your programme came to an error, use the traceback() method. R Favours calling traceback will show all the functions that were invoked prior to the error, also known as the "call stack" in many languages.

3. browser() Function

To launch R's interactive debugger, the browser() function is added to the functions. It will halt function() from running so that you can look at the function's environment. Items can be altered, examine things in the immediate environment and carry on operating while in debug mode.

Notes

4. recover() Function

Unlike the direct statement, the recover() statement acts as an error handler. R prints the entire call stack in recover() and allows you to choose which function browser to enter. then the chosen location is where the debugging session begins.

2.5.2 Analysis

Data analysis is a subset of data analytics; it is a process where the purpose must be made clear, the data must be collected, pre-processed and then it must be analysed (understood, explored insights). The final phase, visualisation, is crucial for helping people comprehend what is going on in the business.

Steps involved in data analysis:

All of these procedures would be part of the data analysis process for the provided issue statement. Analyse, for instance, the items that are being quickly sold out and the specifics of returning customers at a retail store.

- Defining the problem statement – Understand the goal and what is needed to be done. In this case, our problem statement is – "The product is mostly sold out and list of customers who often visit the store."
- Collection of data – Not all the company's data is required; identify the information that is pertinent to the issue. The needed columns in this instance are the product ID, customer ID and visit date.
- Preprocessing – Before undertaking analysis, cleaning the data is necessary to put it in a structured format.
  1. Removing outliers (noisy data).
  2. Removing null or irrelevant values in the columns. (Change null values to mean value of that column.)
  3. If there is any missing data, either ignore the tuple or fill it with a mean value of the column.

2.5.3 Generating Random Numbers

Contrary to pseudo-random numbers, which may offer more security and hacker protection, real random numbers cannot be decrypted using a random seed. True random values are also more akin to nature, which would make them better suited for simulation studies and random experimentation.

To use the random package's functionality, first install and load the package:

install.packages("random")

After installing the package, now load a random library into the R console.

library("random")

2.5.4 Simulating in a Linear Model

A method for creating random data from stochastic processes with known

Notes

parameters is called data simulation. This is something previously done multiple times this semester, albeit not under the heading of "data simulation." For example,

x <- rnorm(100, 3, 0.75)

is a simple data simulation to generate random samples from a normal distribution with known mean (μ=3=3) and variance (σ2=0.752=0.56252=0.752=0.5625).

In recent years, research into machine learning (ML) has grown in tandem with advancements in computing power. As a result, several ML models have undergone significant progress, if not leading to the creation of a brand-new model that outperforms the conventional model.

Lack of the appropriate real-world dataset that complies with the model's assumptions is one of the key issues that researchers typically run into while attempting to execute the proposed model. Or, on the other hand, the real-world dataset may exist but be extremely costly and challenging to gather.

The researchers typically create a simulated dataset that adheres to the model's presumptions to get around these issues. With the simulated dataset being more cost-effective than the real-world dataset, it can be used as a benchmark for the model or to replace it in the modelling process.

The following explanation will describe how a simulated dataset is created. Start by defining the model want to simulate. The coefficients of each independent variable are then calculated and the independent variable and error are then simulated to follow a probability distribution. Finally, depending on the simulated independent variable (and its predetermined coefficient) and error, compute the dependent variable.

2.6 Case Study

2.6.1 Case Study

Let's try scripting some fundamental commands on Palmer Penguins, one of R's most well-known datasets. You should code alongside us to experience R's power.

The Palmer Penguins dataset must first be loaded into the environment.

To install any package in an R environment, we follow the syntax:

install.packages("package name")

This is a single action. You just need to load the package in the environment once it has been installed on your machine to use it. The syntax to load a package is as follows:

library(package name)

Take note that the # symbol denotes a comment, which we insert in the code to make it more readable.

Basic Exploratory Data Analysis

Let's learn more about the Palmer Penguins package now that it is loaded and ready for manipulation. Bear in mind that the dataset in this package is called "penguins".

Viewing the first few rows of any dataset is the traditional way to begin any dataset study. This gives us an understanding of how the dataset is constructed without completely filling the environment with rows. Specifically, we employ the function:

head(name of the dataset)

Notes

A tibble: 6 × 8

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007

The head() function by default only shows the first six rows of the dataset. However, we may provide the number of rows in the brackets like such if we want to examine more/ fewer rows:

head(name of the dataset, number of rows)

A tibble: 10 × 8

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
Adelie	Torgersen	42.0	20.2	190	4250	NA	2007

Similarly, we can also view the last few rows of the dataset using the function:

tail(name of the dataset)

A tibble: 6 × 8

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Chinstrap	Dream	45.7	17.0	195	3650	female	2009
Chinstrap	Dream	55.8	19.8	207	4000	male	2009
Chinstrap	Dream	43.5	18.1	202	3400	female	2009
Chinstrap	Dream	49.6	18.2	193	3775	male	2009
Chinstrap	Dream	50.8	19.0	210	4100	male	2009
Chinstrap	Dream	50.2	18.7	198	3775	female	2009

Notes

The tail() method, like the head() function, shows the last six rows by default. whichever, by adding a parameter, we may modify the method to meet our needs and display whichever many rows we like:

tail(name of the dataset, number of rows)

A tibble: 10 × 8

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
Chinstrap	Dream	50.2	18.8	202	3800	male	2009
Chinstrap	Dream	45.6	19.4	194	3525	female	2009
Chinstrap	Dream	51.9	19.5	206	3950	male	2009
Chinstrap	Dream	46.8	16.5	189	3650	female	2009
Chinstrap	Dream	45.7	17.0	195	3650	female	2009
Chinstrap	Dream	55.8	19.8	207	4000	male	2009
Chinstrap	Dream	43.5	18.1	202	3400	female	2009
Chinstrap	Dream	49.6	18.2	193	3775	male	2009
Chinstrap	Dream	50.8	19.0	210	4100	male	2009
Chinstrap	Dream	50.2	18.7	198	3775	female	2009

To know the dimensions of the dataset, we use the dim() function. The syntax of the function is:

dim(name of the dataset)

344 · 8

Here we can see that our dataset has 344 data points and 8 features, namely- species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, sex and year. Also, we can see that the columns- species, island and sex have the data type factor <fct> (which means categorical data), the columns- bill_length_mm and bill_depth_mm have the data type double <dbl> (which means numbers with decimals) and the columns — flipper_length_mm, body_mass_g and year have the data type int <int>(which means numbers without a decimal point).

A simpler way to know all these details is by using the str() — structure function. The syntax of the str() function is as follows:

str(name of the dataset)

```
tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

Using the str() function, we get to know the dimensions of the dataset, the data type of the columns and the first few values as well.

Here, we also gain deeper understanding, such as the several “levels,” “classes,” or “categories” in the columns with the factor data type, such as the two levels of “sex” (female and male) and the three levels of “species” (Adelie, Chinstrap and Gentoo).

Notes

Summary

- R is a major platform for recently created interactive data analysis techniques. It has expanded quickly and is supported by a sizable number of packages. However, most R programmes are essentially transient and created for a single data analysis task.
- R is frequently used to study and visualise data through graphical presentation and statistical computing. R is a well-known computer language used for graphical display and statistical computation. It is most frequently used to examine and display data.
- An environment and programming language for statistics and statistical graphics is called S.
- R data types are used to define the sorts of data that can be stored in a variable. The proper data type must be chosen for efficient memory usage and accurate computation. Each type of R data has its own set of rules and limitations.
- Extensions known as RStudio add-ins offer a quick method for running complex R functions from within RStudio.
- Tidy data isn't always the simplest to read by humans, it is suitable for sending into other R functions, especially those that are part of the tidyverse.
- Regular Expressions (also known as regex) are a collection of commands that match patterns and are used to find string sequences in enormous amounts of text. These instructions are flexible enough to handle any text or string class because they are made to match a family of text.
- A set of functions called "string manipulation" are used to extract data from text variables.
- A function is simply a section of code that may be called and executed from anywhere in your programme. They are employed to decipher our code in manageable chunks and stay away from repetitive codes.
- Lexical scoping, or static scoping, is used in R. Dynamic scoping is an alternative to lexical scoping that some languages use. Lexical scoping proves to be especially beneficial for streamlining statistical calculations.
- To make things easier, R includes some incredibly helpful functions that implement looping in a condensed manner. Inherently vectorized functions make up the extremely large and robust family of apply functions. These R functions let you apply a certain function to a group of objects, such as files, vectors, matrices, or dataframes.

Glossary

- S: An environment and programming language for statistics and statistical graphics is called S.
- read.delim(): This approach is utilised to read "tab-separated value" (or ".txt") files
- read_tsv(): This method is used for to read a tab separated ("\t") values by using the help of readr package.
- read_lines(): This method is used for the reading line of your own choice whether it's one or two or ten lines at a time.
- read.table(): read.table() reads a file in table format using a general function. Data frames will be used to import the data.

Notes

- read_file(): The entire file can be read using this technique. Must be imported the reader package to use this method.
- read.csv(): read.csv() is used to read “comma separated value” (.csv) files. The data will also be imported in this as a data frame.
- Web scraping: Web scraping is a method for turning unstructured data from HTML tags that are present on the web to a structured format that can be readily accessed and used.
- Tidyverse: The Hadleyverse, a collection of R packages developed or maintained by Hadley Wickham, has a new name: Tidyverse
- Data wrangling: The element of any data analysis job that requires the most time is typically data wrangling

Check your Understanding

1. On which platform R work?
   a) Windows
   b) MAC
   c) Linux
   d) All the above
2. In which year R programming language is developed at the University of Auckland in New Zealand?
   a) 1991
   b) 1995
   c) 1993
   d) 1997
3. What is the formula for ‘removing an object’ from the environment?
   a) >>rm(list=ls())
   b) >>rm(cities)
   c) >>ls()
   d) >>save()
4. Which approach is utilised to read “tab-separated value” (or “.txt”) files?
   a) read_tsv():
   b) read.csv()
   c) read.delim()
   d) read.lines()
5. In which R data type TRUE and FALSE are the only two possible values?
   a) Logical Data Type
   b) Boolean Data Type
   c) Neither a nor b
   d) Both a and b
6. What is/are the different way/s of data scraping?
   a) Human Copy-paste

Notes

b) API Interface

c) DOM Parsing

d) All the above

7. What is the old name of Tidyverse?

a) Hadleyverse

b) Tidyr package

c) Tidydata

d) None of the above

8. How many big packages are there in tidyverse?

a) 5

b) 4

c) 3

d) 6

9. Any value that is contained in quotations ____is a string in R.

a) (‘’)

b) (“”)

c) (,,)

d) ()

10. Which function is used to replace each character in a string?

a) nchar()

b) substr()

c) chartr()

d) None of the above

11. Which function utilised to divide a string depending on a standard?

a) strsplit()

b) setdiff()

c) sub()

d) gsub()

12. The first match in a string can be located and replaced using ___.

a) strsplit()

b) setdiff()

c) sub()

d) gsub()

13. Which function is used returns the matching string’s index or value?

a) grep

b) grepl

c) regexpr

d) regexec

Notes

14. Which function is a hybrid of regexpr and gregexpr?
    a) grep
    b) grepl
    c) regexec
    d) None of the above
15. The element of any data analysis job that requires the most time is typically____.
    a) Scraping
    b) Data Wrangling
    c) Data visualisation
    d) Bata Manipulation
16. A_____ is a set of instructions that are repeated repeatedly until a particular condition is met.
    a) If Statement
    b) If else statement
    c) Loop statement
    d) Nested loop
17. To halt iterations and move control outside of a loop (repeat, for, while), utilise a ____ inside the loop.
    a) break statement
    b) repeat statement
    c) for statement
    d) while statement
18. What is the specifier of full weekday in R programming?
    a) %w
    b) %A
    c) %a
    d) %W
19. What is the specifier of full month in R programming?
    a) %m
    b) %M
    c) %b
    d) %B
20. Which loop function apply a function over subsets of a vector?
    a) sapply()
    b) lappy()
    c) tapply
    d) lappy()

Notes

Exercise

1. Briefly explain the concept of R-Basic.
2. What is the differentiation between R and S programming?
3. What is the installation process of R?
4. What are the add-inns available on R?
5. Describe R controls.
6. Describe R function
7. What is loop function and its type?
8. Explain different type of R tools.

Learning Activities

1. Install R on your system and execute R function.
2. Explain type of loop function with example on R environment.

Check your Understanding-Answers

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

Further Readings and Bibliography

1. Field Cady. The Data Science. 2017
2. William Vance. Data Science: 3 Book in 1 – Beginner's Guide to learn the Realm of Data Science. 2020
3. Peter Bruce Andrew Bruce. Practical Statistics for Data Scientist. 2020
4. Reema Thareja. Data Science and Machine Learning using Python. 2022
5. Uma Maheshwari R Sujatha. Introduction to Data Science: Practical Approach with R and Python. 2021

Module - III: Data Wrangling

Learning Objectives

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

- Understand the basic of data wrangling
- Explain importance of data wrangling
- Analyse the challenges of data wrangling
- Describe web scraping
- Discuss concept of dirty data
- Analyse data manipulation
- Understand bias in data science
- Describe machine learning algorithm

Introduction

Data analytics is fundamentally about manipulation. Of course, don't mean the cunning kind; rather, mean the data kind. Data manipulation is involved in a variety of tasks, including statistical analysis, dashboard creation and data scraping from the web. However, must first make sure that our data are in a usable format before can accomplish any of these things. Here's when the most significant type of data manipulation enters the picture: data manipulation.

The need for accurate data organisation for analysis is growing as the world of data is rising so quickly. Almost all company decisions are made by business users who rely on data and information. Making raw data useable for analytics is therefore crucial. Raw data must be transformed and mapped to be available for analysis, a process known as data wrangling.

3.1 Basics of Data Wrangling

The process of transforming unusable data into a useful form is known as data wrangling. Data munging and data cleanup are some names for it. Prior to performing any data analysis, you should often go through the data wrangling procedure to make sure the data are accurate and comprehensive.

Notes

3.1.1 What is Data Wrangling

Data wrangling, also known to as data munging, is the act of rearranging, changing and mapping data from one “raw” form to another to increase its usefulness and value for a range of downstream uses, such as analytics.

Data wrangling is the process of transforming raw data into the required format and structuring it so that analysts can use it to make decisions quickly. Data wrangling, also refered as data cleaning or data munging, allows businesses to analyse more complex data more quickly, produce more accurate results and draw better conclusions.

3.1.2 Why is Data Wrangling Essential

Some people might wonder whether the time and effort spent on data wrangling is worthwhile. You can comprehend by using a straightforward analogy. Before the above-ground portion of a skyscraper is built, the foundation is expensive and time-consuming. However, this sturdy base is crucial for the structure to stand tall and fulfil its function for many years.

Like data processing, once the infrastructure and code are assembled, it will produce results right away (and perhaps almost quickly) for as long as the process is applicable. However, omitting key data wrangling stages will result in serious drawbacks, missed opportunities and flawed models that harm the organisation’s reputation for analysis.

Data processing has become so dependent on data wrangling tools. The main benefit of utilising data wrangling tools is that:

- Making feasible raw data. The use of quality data in the analysis further down the line is ensured by accurately wrangled data.
- Gathering all information from many sources into one place so it may be utilised.
- Assembling raw data in the necessary manner and comprehending the business context of data.
- Automated data integration tools are employed as data wrangling methods to clean and transform source data into a consistent format that can be applied repeatedly in accordance with end requirements. These standardised data are used by businesses to carry out vital cross-data set analytics.
- Removing noise or inaccurate, missing items from the data.
- Data wrangling serves as a preliminary step in the data mining process, which entails collecting data and analysing it.
- Assisting business users in taking decisive action quickly.

Before data is ready for analytics, data wrangling software typically goes through six iterative steps: discovering, structuring, cleaning, enriching, validating and publishing.

3.1.3 Challenges of Data Wrangling

A critical phase in the data analytics process is data wrangling, which involves converting raw data into a more useful and intelligible form for additional analysis.

There are a few key challenges often face when wrangling data:

1. Scalability and Performance
2. Handling Unstructured and Semi-Structured Data

3. Evolving Data Sources and Formats
4. Privacy and Security Concerns

Notes

3.1.4 Research to Results

Cleaning, manipulating and organising raw data into a format appropriate for analysis is known as data wrangling, sometimes known as data munging or data preparation. Investigating various handling and preparation strategies for data to prepare it for further analysis or modelling is a key component of research that strives to produce successful data wrangling. Here is a description of the data wrangling phase of the research process:

1. Problem Identification: Finding the precise issue or research question is the initial step in any data wrangling-related research. Large datasets, inconsistent data formats, missing numbers, or any other data-related difficulties that need to be resolved could all be examples of this.
2. Literature Review: To grasp the current methods, resources and best practises for data wrangling, researchers frequently undertake a literature review. To learn more about various procedures and approaches that have been applied in the subject, they research academic papers, industry publications and pertinent books.
3. Methodology Selection: Based on the problem identified and the insights gained from the literature review, researchers select appropriate methodologies and techniques for data wrangling. This may involve a combination of manual data cleaning, automated data transformation algorithms, statistical imputation methods for missing values and other data preprocessing techniques.
4. Experimental Design: Researchers design experiments or studies to evaluate the effectiveness of the selected methodologies. This could involve creating controlled datasets with known issues, simulating specific data problems, or working with real-world datasets to validate the chosen techniques.
5. Data Collection: Researchers gather pertinent datasets for their experiments based on the study's goals. Data from numerous sources, such as databases, APIs, online repositories, or data created especially for the study, could be included.
6. Data Cleaning: To get rid of errors, duplication, inconsistencies and outliers, the acquired data frequently has to be cleaned. To find and fix these problems, researchers can use custom scripts, employ data cleaning technologies, or use manual processes.
7. Data Transformation: The data is transformed using a variety of approaches by researchers so that it may be analysed. This could entail altering data, changing variables, compiling data, or developing derived characteristics based on domain expertise.
8. Missing Data Handling: Missing values are a frequent problem in datasets from the real world. To successfully manage missing data, researchers investigate various imputation strategies such as mean imputation, regression imputation, or multiple imputation.
9. Quality Assessment: To make sure the processed data is in line with the goals of the study, researchers assess its quality and integrity. In this step, data consistency, accuracy and completeness are evaluated. Any outstanding problems are also noted.
10. Documentation: It is essential to fully document the data wrangling procedure. Researchers provide transparency and reproducibility by documenting the actions completed, judgements made and any changes or imputations made.

Notes

11. Result Analysis: The output of the data wrangling process is examined by researchers, who then interpret the findings. To acquire insights into the data and determine whether the research objectives have been reached, this analysis may include exploratory data analysis, statistical summaries, or visualisation approaches.
12. Conclusion and Recommendations: Researchers make conclusions and suggestions for enhancing data wrangling methods or resolving issues raised during the research based on the findings.

Researchers hope to create useful data wrangling methodologies, algorithms and guidelines through the research process described above that may be used in a variety of domains and datasets. These results support more precise and trustworthy data analysis and modelling practises and advance data preprocessing techniques.

3.2 Data Refinement

A data boom is happening across the planet. Every day, produce up to 2.5 quintillion bytes of data and that number is growing. Customers', partners' and ERP systems' generated data has the potential to give your business a significant competitive advantage. This data needs to be prepared, compiled and properly assessed, which is known as data refinement. Data today must be treated more like manure than like gold.

"Data is not so much like gold as it is more like manure. Having a big pile of it does nothing, but you need to know how to spread it around to make your business grow."

The process of data refining is essential to creating a data-driven business and upholding ethical practises.

3.2.1 Wrangling (Data Import Considerations)

The six-step data wrangling procedure, which covers everything needed to make raw data usable, is described here.

- Step 1: Data Discovery
- Step 2: Data Structuring
- Step 3: Data Cleaning
- Step 4: Data Enriching
- Step 5: Data Validating
- Step 6: Data Publishing

Notes

Step 1: Data Discovery

The process of Data Wrangling starts with discovery. This is a common way of saying that you have understood or are familiar with your data. You must think about how you would like to organise the data in order to make it simpler to consume and analyse.

Step 2: Data Structuring

When raw data is gathered, it comes in a variety of sizes and forms. It lacks a clear structure, which indicates that it lacks a model and is wholly disorganised. Giving it a structure makes it easier to rearrange it to suit with the analytical model that your company has implemented and makes for better analysis.

Unstructured data frequently has a lot of text and contains elements like dates, numbers, ID codes, etc. The dataset must be parsed at this stage of the Data Wrangling procedure. Relevant information is taken from new data using this technique. For instance, if you are working with HTML code that was scraped from a website, you might parse it to extract the information you need and reject the rest.

As a result, a spreadsheet with meaningful data and columns, classes, headings and other features will be more user-friendly.

Step 3: Data Cleaning

Data wrangling and data cleaning are terms that are frequently used interchangeably. Nevertheless, these are two completely distinct processes. Cleaning is only one part of the larger process of Data Wrangling, albeit being a difficult procedure.

Cleaning the data does the following:

- It eliminates outliers from your dataset that can cause your data analysis results to be skewed.
- It standardises the data format and changes any null values to enhance quality and consistency.
- It locates duplicate values, standardises measurement systems, corrects grammatical and typographical problems and validates the data to make it easier to handle.

Several technologies, including Python and R, can be used to automate certain algorithmic processes.

Step 4: Data Enriching

You have a thorough comprehension of the data available to you at this point in the Data Wrangling process and have grown accustomed to it.

You only need to perform the optional step of enriching the data if the current data doesn't satisfy your criteria.

Step 5: Data Validating

To resolve any difficulties with the quality of your data with the right transformations, validating the data is a necessary action.

The rules of data validation require repetitive programming processes that help to verify the following:

- Quality

Notes

- ❖ Consistency
- ❖ Accuracy
- ❖ Security
- ❖ Authenticity

This is an excellent illustration of the overlap between data cleaning and data wrangling, which can occasionally occur.

You may need to perform this procedure numerous times because errors are likely to be discovered.

Step 6: Data Publishing

All the steps have been finished by this point and the data is prepared for analyses. The newly wrangled data must now be published in a location where it can be easily accessed and used by you and other stakeholders.

The information can be added to a fresh architecture or database. The result of your work will be high-quality data that you can utilise to obtain insights, produce business reports and more if the other steps were appropriately carried out. Even more data processing could be done to produce more elaborate and substantial data structures, like data warehouses. The possibilities are unlimited at this point.

3.2.2 Web Scraping

Web scraping is the practise of extracting data from websites. Many websites don't let users keep information for personal use while they are online browsing. The data can be manually copied and pasted, but this process is labour- and time-intensive. The process of deleting information from websites is automated via web scraping.

Uses of Web Scraping:

Web scraping has a variety of uses, both professionally and personally. Some common applications of web scraping are as follows, each with varying needs at various levels.

- Brand Monitoring and Competition Analysis: Web scraping is employed to gather consumer reviews of a particular service or good to ascertain how the customer thinks about it. In a structured, useable format, competitor data is also extracted using it.
- Machine Learning: Artificial intelligence is used in machine learning, which relies on experience rather than explicit programming to help the machine learn and grow. Millions of websites' worth of data are needed for that and web scraping software is used to retrieve it.
- Financial Data Analysis: Web scraping is used to retain a usable record of the stock market, so it can be used for insights.
- Social Media Analysis: It is used to gather information from social media platforms on customer patterns and their responses to campaigns.
- SEO monitoring: Search engine optimisation is the process of enhancing a website's visibility and rating across a variety of search engines, including Google, Yahoo, Bing and others. To comprehend how the ranking of the material changes over time, web scraping is performed.

Notes

Techniques of Web Scraping:

There are two methods for obtaining information from websites: manual extraction and automated extraction.

- Manual Extraction Techniques: The manual copying and pasting of the website's content falls under this category. Although tedious, time-consuming and repetitive, it is a successful method for obtaining data from websites with powerful anti-scraping safeguards, such as bot detection.
- Automated Extraction Techniques: According to user needs, web scraping software is used to automatically scrape data from websites.
- HTML Parsing: Making anything comprehensible for component-by-component analysis is referred to as parsing. To put it another way, it implies converting information from one form to another that is simpler to work with. HTML parsing entails reading the code and pulling out the pertinent information based on what the user needs. The primary platform for execution is JavaScript and the intended endpoint is HTML pages.
- DOM Parsing: The World Wide Web Consortium's official proposal is the Document Object Model. It outlines an interface that a user can use to edit the XML document's style, structure and content.
- Web Scraping Software: Many online scraping technologies are available now or have been developed specifically to meet user needs for obtaining desired information from millions of websites.

Tool for Web Scraping:

Web scraping software is designed specifically to collect data from the internet. They are sometimes referred to as web harvesting tools or data extraction tools and they are helpful for anyone attempting to gather specific data from websites since they give the user structured data by pulling information from a number of websites. Some of the most popular Web Scraping tools are:

- ❖ Import.io
- ❖ Webhose.io
- ❖ Dexi.io
- ❖ Scrapinghub
- ❖ Parsehub

Legalization of Web Scraping:

It's a touchy subject, but legalising online scraping might either be beneficial or detrimental, depending on how it's use. site scraping using a reliable bot, on the other hand, helps search engines to index site material and price comparison services to help customers save money and get the best deal.

Web scraping, however, can be redirected to serve more nefarious and harmful purposes. Web scraping is sometimes associated with other hostile automation techniques known as "bad bots," which facilitate destructive actions like denial-of-service assaults, competitive data mining, account takeover, data theft, etc. Web scraping's legality is a murky field that seems to grow with time.

Web scraping is the main cause of the rise in copyright violations, terms of service

Notes

violations and other activities that are seriously detrimental to a company's operations, even while it technically speeds up data surfing, loading, copying and pasting.

Challenges to Web Scraping:

There are other issues that pose a hurdle to web scraping in addition to the question of its legality.

- Data Warehousing: A vast amount of information must be kept because of data extraction at scale. The searching, storing and exporting of this data will become laborious tasks if the data warehousing architecture is not adequately established. Therefore, a faultless data warehousing system is required for large-scale data extraction.
- Website Structure Changes: Every website regularly modifies its user interface to enhance usability and appeal. This calls for numerous structural adjustments as well. Web scrapers also need modifications because they were put up using the website's code at the time.

Because inaccurate information about the website's structure will result in improper data scraping, they also need weekly modifications to target the right websites.

- Anti-Scraping Technologies: Some websites employ anti-scraping technologies to prevent any attempt at scraping. To stop any bot interference, they employ a dynamic coding method and an IP blocking system. Working around such anti-scraping technology takes a lot of time and money.
- Quality of Data Extracted: The overall integrity of the data will be impacted by records that do not include the required level of information. It is challenging to ensure that the data scraped adheres to the quality standards because it must be done in real-time.

Future of Data Scraping:

It may be properly said that unintentional data-scraping practitioners are likely to generate a moral hazard when they target the companies and retrieve their data because there are some problems and opportunities for it.

Data-scraping in combination with big data, however, can give the organisation market information, assist them uncover important trends and patterns and identify the greatest prospects and solutions because we are on the cusp of a data transformation. Therefore, it would be accurate to suggest that data scraping will soon be improved.

3.2.3 Types of Dirty Data

The one bad fruit that can ruin your entire marketing and sales strategy is dirty data. Learn about the 8 different types and how to clean your CRM. One of the biggest sources of corporate expense is dirty data, which costs an astounding $13 million year on average. And the effects go beyond the financial ones.

Your reps have trouble tracing the lead's origin with filthy data. They waste valuable time and suffer from poor productivity. In addition to this, it also results in resource disruptions, ineffective internal and external communication and lost marketing expenses.

In contrast, a strong revenue operations function depends on high-quality data. Leaders can get timely and useful insights, streamline procedures and make wise company decisions with the aid of accessible and pertinent data.

Notes

Understanding what dirty data is, how it affects business and how to deal with it are crucial given that it has a terrible impact on it (a frightening understatement).

What is "Dirty" Data?

Every organisation needs data, but not all of it has value for your company. Dirty data is the one bad apple that destroys the entire marketing and sales basket.

Up to 74% of organisations acknowledge that better data management is necessary in 2022 to prevent financial and competitive disadvantages.

Dirty data, in essence, is erroneous information that interferes with a company's database and has an impact on critical operations like GTM, segmentation, customization, lead scoring, prospecting and planning for optimal customer profiles, among others. The outcome? Ineffective business practises, missing chances, inefficiencies and harm to reputation.

And that isn't all. Poor data quality costs businesses an average of $15 million year in losses, which is where it harms them the most.

Mostly through manual data entry, human mistake, ineffective departmental communication, or third-party integrations, dirty data enters the CRM. You must comprehend the many forms of unclean data and how to clean it if you want to ensure that every lead touchpoint is special.

Types of Dirty Data (and How to Clean It)

Although filthy data can take many different forms, we have grouped it into 8 types. Take a look below.

1. Duplicate Data

The most prevalent issue with data quality is data duplication. Only a few data points in the CRM are accidentally shared with other records, such repetitive leads, accounts and contacts. While carbon copy duplicates are the easiest to spot and eliminate, partial copies, which occasionally arise due to human error, present more serious problems.

Duplicate data can result in inaccurate data recovery, ineffective personalisation, wasteful workflows, overcrowded storage systems and repetitive customer interactions.

For instance, when it comes to ABM, each account receives or anticipates a customised encounter. The prospect might think your campaign is automated and impersonal if you have the same prospect listed in your database three times and send them the identical email repeatedly. It just serves to irritate the potential customer, decreasing conversion possibilities.

How to clean Duplicate Data?

Manual data cleansing is insufficient in the present environment, when firms deal with massive amounts of data every day. Additionally, partial duplicates are not always eliminated by hand cleaning. Instead, you can spend money on an automation platform that finds and eliminates duplicate data and merges similar data. Additionally, it may combine duplicate data using criteria specific to your business and categorise it.

2. Insecure Data

Security restrictions have changed the marketing landscape, driven by the growth of data. Parallel to this, severe privacy problems have harmed consumer-firm interactions,

Notes

causing legislative interventions as well as changes in people's privacy-protective behaviours.

The GDPR and CCPA are two significant privacy and data security legislation that are now in effect. Data that doesn't adhere to these standards or isn't secure may be subject to severe financial penalties. A user can have previously submitted data without agreeing to your data sharing and privacy policy, for example. These types of unsecure data can have detrimental effects.

Consumers are becoming more and more the focus of company decisions nowadays and digital permission, opt-ins and privacy notifications are evolving into the new norm. Without strong CRM hygiene, compliance with these regulations becomes nearly impossible.

Not to mention the detrimental effect on brand reputation. Despite facing backlash from the public, companies like Amazon and WhatsApp have already paid substantial fines totalling more than $800 million and $270 million, respectively, for suspected GDPR non-compliance.

How to clean Insecure Data?

Compliance with data privacy rules can be directly aided by maintaining a clean database. Best practises for cleaning insecure data include eliminating useless and unsafe CRM entries, merging duplicates for more current data, simplifying your data stack, automating the lead-to-account connecting process and hosting your CRM on legally sound cloud software.

3. Outdated Data

A piece of information that seems important today could not be useful tomorrow. Analytics based on out-of-date data are equivalent to using the wrong GPS while driving and going over a precipice.

Think about this. To obtain your resource, a website visitor must fill out a form. In the months that follow, they become a prospect and engage with your business more by subscribing to newsletters and replying to emails. However, this information hasn't been changed in your CRM.

Because of this, the information you give them is still intended for a brand-new lead rather than one who is already being cultivated. It prevents them from moving further down the conversion process to become consumers.

Aside from employment changes, organisational restructuring or mergers and old software systems that can't keep up with the quick pace of technological innovation, other causes of outdated data could be.

How to clean Outdated Data?

Prior to migration or system integration, data should be purged and cleansed to get rid of stale information. Identifying your company's key moment is another important step. Delete all historical data from the system. Automation can do this process for you in a matter of hours, but manual cleaning can take days or weeks. So, use an automated tool instead.

4. Incomplete Data

A record is deemed incomplete if it lacks critical components needed to process incoming data before sales and marketing act. Sales salespeople' jobs are made much

Notes

more difficult by data shortages. 45% of sales representatives struggle with knowledge gaps. Unfortunately, problems with incomplete data are frequent, which makes it more difficult to maximise the value of the data you've gathered.

Take the example that you have the customer's phone number but not their email address. You miss out on a big sales opportunity since you can't send this customer a key email campaign promoting your products.

In addition, it is difficult to score leads and segment prospects based on insufficient information.

How to clean Incomplete Data?

You are given two choices. The first step is to perform a manual search and complete incomplete records by adding missing data. However, you'll soon realise that this strategy is neither workable nor scalable. In contrast, automated lead capture automatically completes forms for you and provides thorough account details.

5. Inaccurate Data

One of the worst types of data pollution is inaccurate data. Reps may have filled out a field accurately, but the data is fictitious or incorrect.

Consider a scenario when a potential customer provides a bogus mobile number, preventing your representative from contacting them via phone. Even worse, when accuracy is crucial, communicating with the incorrect individual might cause the entire purchasing process to be disrupted.

A huge 77% of companies think that having erroneous data makes it harder for them to adjust with the times. Additionally, 41% of sales representative's report having trouble handling erroneous data. It not only results in inaccurate reporting and bad judgement, but it also delays the development of an opportunity.

How to clean Inaccurate Data?

Starting from the beginning, the data used for marketing and sales initiatives must be accurate. It is crucial to monitor this data at its point of entry and prevent it from entering the system. Integrate with a platform for automated data collection like Nektar to improve accuracy.

6. Incorrect Data

Information that is incorrectly stored or information that doesn't meet the requirements of a particular field. For instance, a designation may be under the field for the firm name, or a text field could include a number value. This faulty data has several serious drawbacks, including poor campaign targeting, irrelevant communication and a dearth of prospect insights.

39% of organisations reported negative effects on experience owing to poor data quality in a B2B business ecosystem, where improving the buyer experience is a top concern. To deliver the experience that customers want, you should either remove or update this type of data.

How to clean Incorrect Data?

To guarantee that data is accurate or legitimate, reps should adhere to standards and enter information within the permitted ranges. Additionally, you can programmatically enforce the accuracy of data points by using lookup tables or edit checks.

Notes

7. Inconsistent Data

Don't confuse duplicate and inconsistent data even though they look to be the same. Data is duplicated when it is copied exactly as is. Unreliable data, on the other hand, deviates from standards and doesn't follow predefined guidelines.

When the same element appears in different versions throughout the system, inconsistent data is shown. For instance, the identical data field is used for the Chief Marketing Officer, Chief of Marketing, Chief Marketing Officer and Chief Mktg Officer, but they are input in different ways.

Since sales representatives must consider many variables of the same lead information, inconsistent data has a negative impact on analytics and decision-making.

How to clean Inconsistent Data?

The centralised technique is the quickest way to clear this dirty data. Your salespeople can adhere strictly to a uniform file naming convention that you establish. Utilise an automatic data collection programme for CRM to handle the heavy lifting for you instead of having to manually remove any existing incorrect data, which may be challenging.

8. Hoarded Data

Data hoarding can impede data exchange and increase storage costs. Additionally, it results in data hygiene issues, making it difficult to extract critical insights for business decision-making.

Storing an excessive amount of data will increase storage costs and hinder data sharing. Additionally, it leads to issues with data cleanliness, which makes it difficult to extract essential insights for business decision-making.

Occasionally, various departments require various data variables. Because departments can't locate crucial data points in another team's storage, it has a detrimental impact on cooperation while increasing data storage.

How to clean Hoarded Data?

By concentrating on the most important data and making it available in one location, businesses can prevent data hoarding. It shortens the time needed to conduct analyses, enhancing teamwork. Use automated data capturing tools like Nektar, which can simultaneously clear old and accumulated data. It's all about putting time aside.

3.2.4 Manual Dirty Data Forms

It can be simple to overlook the reality that filthy data encompasses more than just inaccurate data in the information-oversaturated world of today. The same way that plastic bags and vehicle emissions harm our environment, filthy data harms our businesses. For businesses of all sizes and in all sectors, dirty data inside CRMs like Salesforce is a major issue.

If you were to Google "dirty data," you would unavoidably find publications that describe duplicated, inaccurate, incomplete and inconsistent data as being "dirty data." But the reality goes beyond that.

The Data Warehousing Institute (TDWI) estimates that firms in the United States lose more than $600 billion annually because of filthy data.

Notes

Unfortunately, maintaining a database is more difficult than maintaining our homes or neighbourhoods. First define what exactly constitutes filthy data to address the data problem. Let's discusses the five categories of unclean data that make up many databases and the methods you can use to clean them up.

1. Duplicate Data

Records or entries with duplicate data unintentionally exchange information with another record in your database. A complete carbon copy of another record is the most typical type of duplicate data. These types of data contamination are the worst. Contacts, leads and accounts are the objects that are most frequently replicated.

2. Outdated Data

Imagine finding a report that is appropriate for your project only to learn later that it is out of date. Therefore, outdated data is essentially information that is false, deficient, or just not in use anymore.

3. Incomplete Data

The most typical type of dirty data is incomplete data. a record that is missing important fields from master data records that are important to business, such as industry type, title, or last names. For instance, you cannot target your sales and marketing initiatives by industry if you neglected to categorise your consumers by industry. Imagine approaching a potential client who is located at "N/A" and trying to offer them geolocation software.

4. Inaccurate/Incorrect Data

In order to better understand your clients and satisfy them, it is helpful to gather information about them. This is only achievable if data is gathered correctly, thoroughly and accurately. It can also result in expensive errors.

a) Incorrect data: When field values are generated outside of the acceptable range of values, it happens. For instance, a month field should only accept values between 1 and 12 and a house or office address must be a genuine address.

b) Inaccurate data: The data in a field is frequently accurate but incorrect when taken into account of the business context. Inexact data can cause expensive interruptions. For instance, even though the address on which it was delivered was valid, mistakes in a customer's address may cause the product to be delivered to the incorrect place.

Stats related to Inaccurate/Incorrect Data:

- 43% of sales and marketing teams think that dealing with a lack of reliable data is difficult for them.

- 54% of B2B companies claim that poor data quality prevents them from becoming successful.

- Unreliable data, according to 69% of Fortune 500 organisations, hinders their operations.

5. Inconsistent Data

When the same field value is saved many times, the result is inconsistent data,

Notes

sometimes referred to as data redundancy. For instance, businesses have client information on numerous systems, but the data is not synced.

If you wish to target all "Vice President" for an upcoming email marketing campaign, you can see how inconsistent data a problem is. However, as "V.P," "v.p," "VP," and "Vice Pres" all signify the same thing, they would only be included in the campaign if all of these variations are listed. When you must take into account all the factors of the same title, industry, etc., inconsistent data complicates analytics and makes segmentation challenging.

Best Practices for Data Cleaning

Following are some of the best practices which can be considered while data cleaning.

1. Create a Data Quality Plan: Setting clear expectations for what an ideal database should look like is essential. It is recommended that you develop KPIs (key performance indicators) for each employee working on your project. How will your workers achieve these KPIs and what are they? Which techniques ought to be applied to account for the data's health? How can data hygiene be consistently maintained?

   You can learn more about error occurrence, recognise wrong data and appreciate the root of data health issues by routinely implementing best practises for data cleansing. This will result in the upkeep and cleaning of data in the future.

2. Standardize Contact Data at the Point of Entry: When you permit incorrect or unhealthy data to into your database, it is very challenging to maintain excellent data hygiene. It is essential to validate data at all points of entry even before the date cleaning may take place. This will guarantee standardised data entry and assist in removing duplicate data.

   One approach to achieve this is to request that your team develop a data entering SOP (Standard Operating Procedure). Only high-quality data will be able to enter your database at the time of entry if you adhere to this SOP.

3. Validate Data Accuracy: Real-time data accuracy validation is difficult. There are some tools for cleaning data, like list imports. Numerous hygiene tools exist, such as address, phone and email verification.

6. Merge Duplicates

Instead of eliminating the duplicate data, try combining it. Because every single piece of data has significance, merging is always advised. However, you must establish a master rule set to make sure the duplicates are merged with the correct contact.

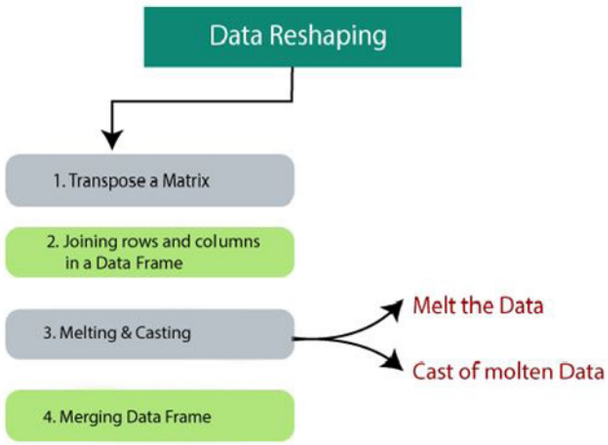
By doing this, you can create fresh data that will automatically match and merge with the master or original record. For instance, if you have five records in Salesforce, you are probably going to use all of the current titles and phone numbers columns from all of the most recent additions while keeping the lead source from the original/master record.

3.2.5 Reshaping Data

Data Reshaping in R is the process of altering the data's row and column arrangements. When processing data in R, a data frame is used as the input. A data frame's rows and columns make it much simpler to extract data, but there is a difficulty

Notes

if need the data frame in a format other than the one given. In a data frame, R has numerous functions to merge, split and transform rows to columns and vice versa.



Data reshaping stages

Transpose a Matrix

By providing the t() function, R enables us to determine the transpose of a matrix or data frame. This t() function accepts a matrix or data frame as an input and returns the matrix or data frame's transposition. The t() method has the following syntax:

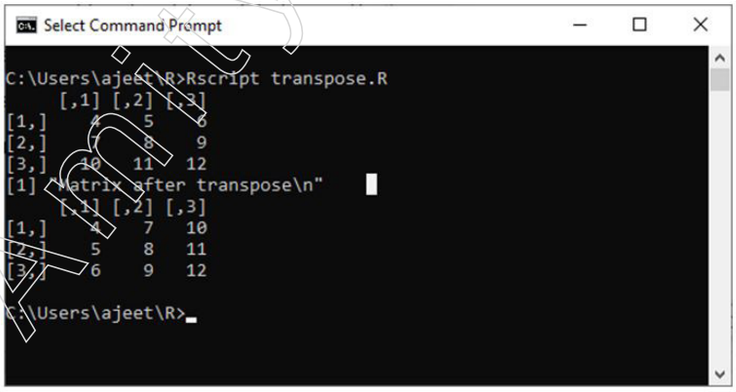
t(Matrix/data frame)

Let's see an example to understand how this function is used

Example

```
a <- matrix(c(4:12),nrow=3,byrow=TRUE)
a
print("Matrix after transpose\n")
b <- t(a)
b
```

Output:



Joining rows and columns in Data Frame

Can connect multiple vectors in R to produce data frames. R offers a method called cbind() for this purpose. Additionally, R has the rbind() function, which enables us to

Notes

combine two data frames. To access information that depends on both data frames, need to merge data frames. The cbind() and rbind() functions have the following syntax.

```
cbind(vector1, vector2,.....vectorN)
rbind(dataframe1, dataframe2,.....dataframeN)
```

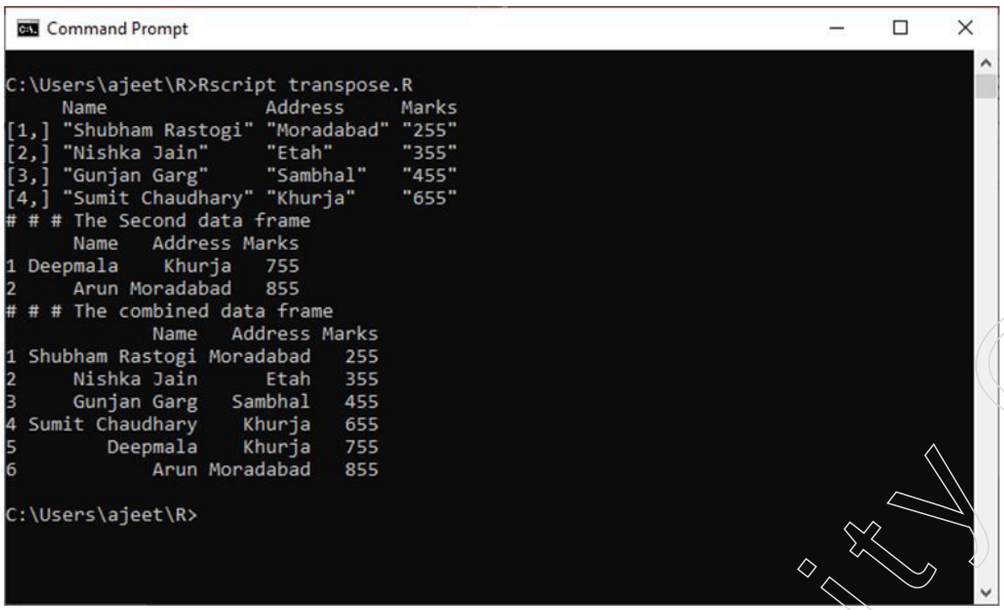
Let's see an example to understand how cbind() and rbind() function is used.

Example

```
#Creating vector objects
Name <- c("Shubham Rastogi","Nishka Jain","Gunjan Garg","Sumit Chaudhary")
Address <- c("Moradabad","Etah","Sambhal","Khurja")
Marks <- c(255,355,455,655)

#Combining vectors into one data frame
info <- cbind(Name,Address,Marks)

#Printing data frame
print(info)

 # Creating another data frame with similar columns
 new.stuinfo <- data.frame(
 Name = c("Deepmala","Arun"),
 Address = c("Khurja","Moradabad"),
 Marks = c("755","855"),
 stringsAsFactors=FALSE
 )

 #Printing a header.
 cat("# # # The Second data frame\n")

 #Printing the data frame.
 print(new.stuinfo)

 # Combining rows form both the data frames.
 all.info <- rbind(info,new.stuinfo)

 # Printing a header.
 cat("# # # The combined data frame\n")

 # Printing the result.
print(all.info)
```

Output:

Notes



Merging Data Frame

The merge() function in R allows you to combine two data frames. There is a requirement for data frames to have the same column names during the merging process.

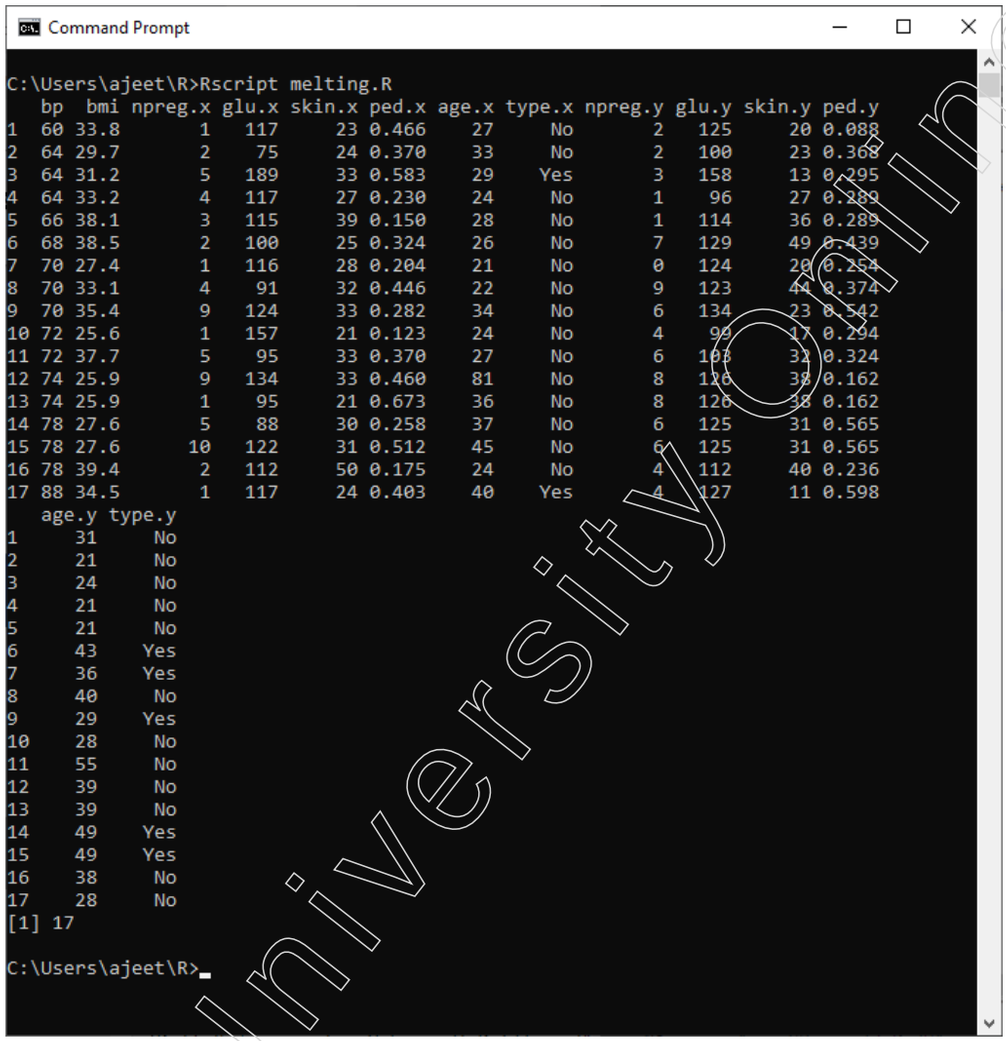
Let's use a dataset from the "MASS" library that describes diabetes in Pima Indian women as an example. Based on the results of the blood pressure and body mass index measurements, combine two datasets. The records where the values of these two variables match in both data sets are joined to create a single data frame when these two columns are chosen for merging.

Example

```
library(MASS)
merging_pima<- merge(x = Pima.te, y = Pima.tr,
by.x = c("bp", "bmi"),
by.y = c("bp", "bmi")
)
print(merging_pima)
nrow(merging_pima)
```

Output:

Notes



Melting and Casting

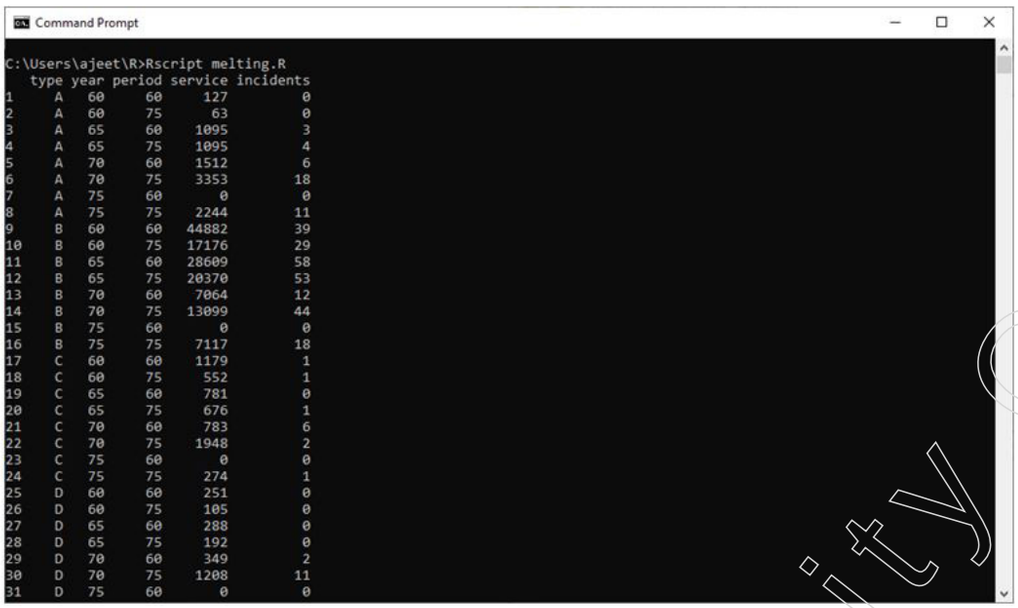
The most significant and fascinating topic in R is how to modify the shape of the data in several steps to get the desired shape. R offers the melt() and cast() functions for this purpose. Consider a dataset named ships from the MASS library to comprehend how it works.

Exampl

```
library(MASS)
print(ships)
```

Output:

Notes



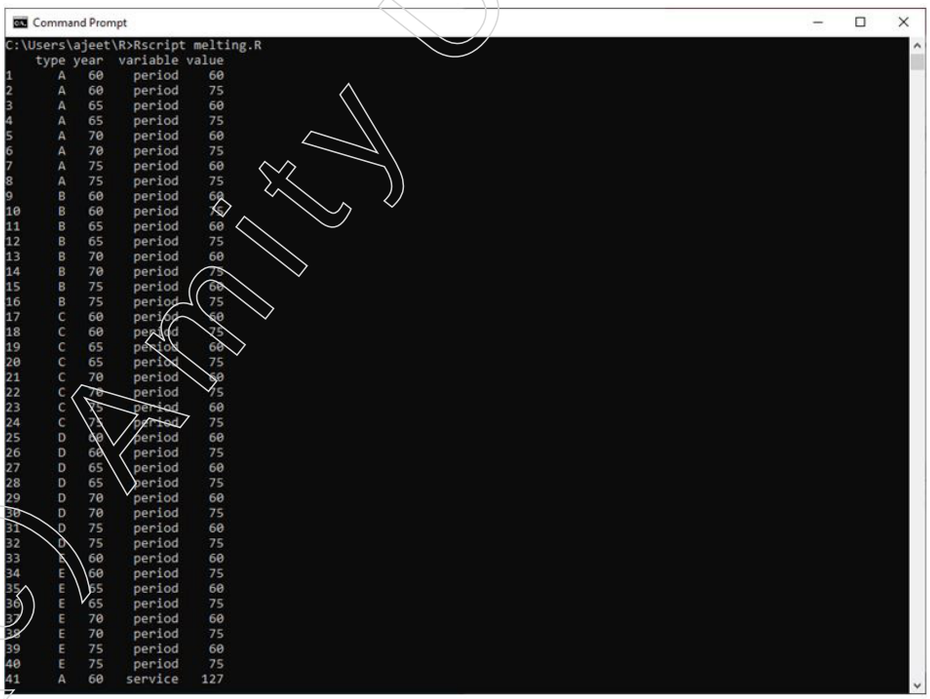
Melt the Data

Now organise the data by melting it. Columns are “melted” when they become several rows. Except for type and year, make the dataset’s columns into numerous rows.

Example

```
library(MASS)
library(reshape2)
molten_ships <- melt(ships, id = c(“type”,”year”))
print(molten_ships)
```

Output:



Notes

Casting of Molten Data

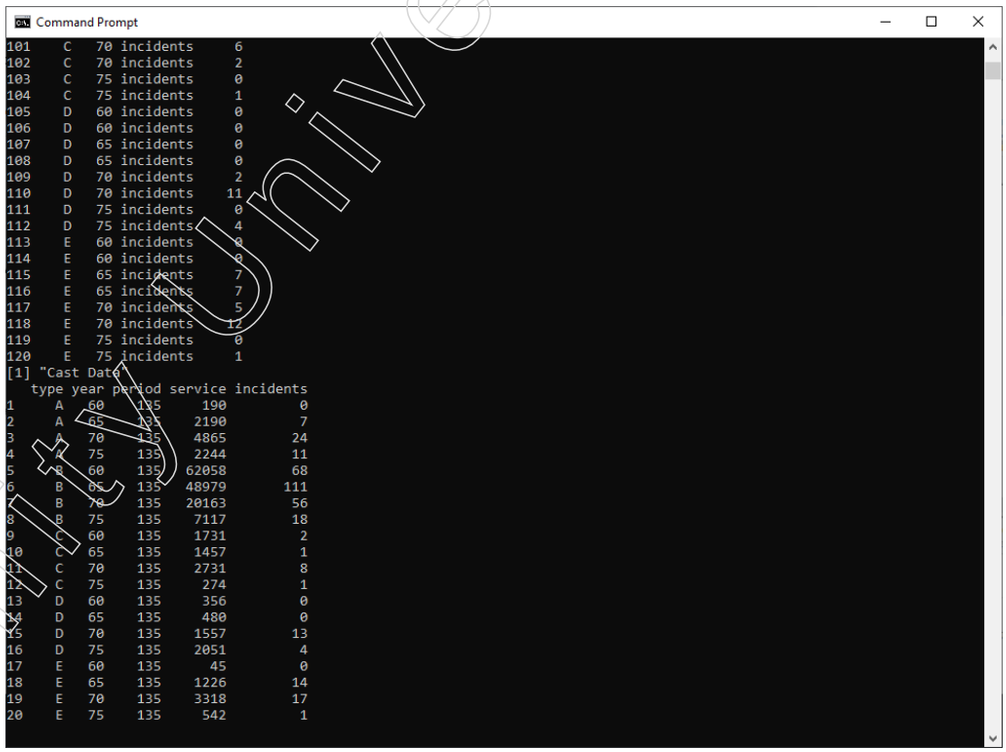
The data can be melted and then cast into a new form to get the aggregate of each type of ship for each year. R has the cast() function for this.

Let's starts doing the casting of our molten data.

Example

```
library(MASS)
library(reshape2)
#Melting the data
molten.ships <- melt(ships, id = c("type","year"))
print("Molted Data")
print(molten.ships)
#Casting of data
recasted.ship <- dcast(molten.ships, type+year~variable,sum)
print("Cast Data")
print(recasted.ship)
```

Output:



3.2.6 Inference or Statistical Inference

Statistical inference is the process of extrapolating features of a population from a sample of data. Using statistical methods and sample data, the characteristics of the complete population from which the sample was drawn are estimated.

Notes

Typically, scientists are interested in studying a population. knowing the findings at a population level is far more beneficial than knowing merely the relatively few study participants when researching a phenomenon, such as the impacts of a new drug or public opinion.

Unfortunately, populations are frequently too big to accurately estimate. Therefore, to learn about it, researchers must use a controllable portion of that population.

You can calculate the characteristics and dynamics of a population by employing techniques that allow for statistical inference. More specifically, population parameters can be estimated using sample statistics. Learn more about the distinctions between population parameters and sample statistics.

Consider that you are researching a new drug, for example. You want to know how the medication affects the entire population, not just a tiny sample, as a scientist. After all, the greater society can't really benefit from knowing the impact on a small number of people.

As a result, you are interested in drawing a statistical conclusion about the impact of the medication on the general population.

How to Make Statistical Inferences

As a result, you are interested in drawing a statistical conclusion about the impact of the medication on the general population:

1. Create a sample that fairly depicts the entire population.
2. Calculate your interest-related variables.
3. While accounting for sampling error, generalise your sample results to the population using the proper statistical methodology.

That is, of course, the simplest explanation. Creating treatment and control groups, giving out treatments and minimising other sources of variance may all be necessary in real-world experiments. In more intricate circumstances, it could be necessary to build a model of a process. A lot of specifics are needed when drawing a statistical inference. Know how to use statistical inference in scientific research.

You must employ a method that allows for statistical inference after getting a representative sample. Even if your sample resembles the population, it will never be an exact replica of it.

Sampling error is the term used by statisticians to describe the discrepancies between a sample and the population. Any association or impact you observe in your sample can simply be sampling error, not a genuine discovery. Results from inferential statistics include sampling error. study up on sampling error.

Notes

Common Inferential Methods

The following are four generally accepted methods for drawing conclusions from statistical data.

1. Hypothesis Testing: Evaluates two opposing population theories using representative samples. Results that are statistically significant after controlling for sampling error suggest that the sample effect or relationship is present in the population.
2. Confidence Intervals: A set of values that most likely includes the population value. By calculating a margin around the estimate and evaluating the sampling error, this approach gives an indication of how inaccurate the estimate may be.
3. Margin of Error: Like a confidence interval, but typically for survey results.
4. Regression Modelling: a projection of the population's outcome-producing process.

3.3 Data Manipulations

The act of arranging data to improve its readability, aesthetic appeal, or structure is known as data manipulation. For instance, a collection could be organised alphabetically to make data of any kind easier to grasp.

Website owners can monitor their most popular pages and traffic sources by manipulating data. It is therefore widely used on web server logs.

Accounting professionals and those in related professions also manipulate data to arrange it to determine things like product costs, upcoming tax liabilities, pricing trends, etc. It also aids stock market forecasters in making predictions about future changes and stock performance.

Additionally, computers may manipulate data to display information to people in a more realistic manner that is based on web pages, software programme code, or data formatting.

The computer language DML is used to manipulate data. It stands for Data Manipulation Language and is used to change databases and add, remove and edit data. It entails modifying the text such that it is easier to read.

Objective of Data Manipulation

A crucial component of company operations and optimisation is data manipulation. You must handle data properly and transform it into useful information using techniques like trend analysis, financial data and customer behaviour. Data manipulation has several benefits for an organisation, some of which are listed here:

- Consistent data: Data manipulation offers a technique to organise your data such that it is organised, easily readable and more easily understood. You might not have a unified perspective of the data when you acquire it from many sources, but data manipulation gives you assurance that the data is properly organised, structured and stored consistently.
- Project data: Data manipulation is more helpful, particularly when it comes to finances because it enables more thorough research by analysing past data to forecast the future.
- Delete or neglect redundant data: Data manipulation assists with data maintenance and permanent deletion of useless data.

Notes

- In general, you can perform a wide range of activities with the data, including edit, remove, update, convert and incorporate data into a database. It aids in maximising the value of the data. Data loses its value if you don't know how to use it effectively. Therefore, when you can organise your data appropriately, it will be advantageous to make better business decisions.

3.3.1 Probability

Probability is a key subject that new data scientists must understand. Probability is an intuitive idea, to put it simply. Daily, employ it without necessarily being aware that we are speaking and using probability in our work.

There are numerous unknowns in life. Until something happens, cannot predict how a situation will turn out. Today, will it rain? Will one succeed on his/her next maths exam? Who will win the coin toss—the preferred team? Will one be promoted in the upcoming six months? These queries are all illustrations of the unpredictability of the world we live in. Let's translate them into a few basic terms that will be use moving forward.

- Experiment – are the ambiguous circumstances with potential for various outcomes. It's an experiment to see if it rains every day.
- Outcome is the outcome of one trial. The result of today's trial from the experiment will therefore be "It rained" if it rains today.
- Event is one or more of an experiment's results. One of the potential outcomes for this experiment is "it rained.".
- Probability is a way to quantify the likelihood of an event. The probability of the outcome "it rained" for tomorrow is 0.6 if there is a 60% chance that it will rain tomorrow.

Why do probability is needed?

The study of random phenomena is the main emphasis of the mathematical field of probability theory. For data scientists working with data that is subject to chance, it is a crucial skill.

The application of probability theory enables the investigation of chance events since randomness exists everywhere. The objective is to estimate the probability of an event occurring, frequently using a numerical scale between 0 and 1, with "0" denoting impossibility and "1" denoting certainty.

A coin flip with two possible outcomes—heads or tails—is a classic illustration of this. Here, there is a 50% chance that a single toss will result in a head or a tail. You might discover that the results of your individual experiment can vary. But as you keep flipping the coin, the result becomes more likely to be 50/50.

In many branches of science, probability is crucial. As a means of summarising their findings, researchers can incorporate uncertainty into their research models. This enables a forecast distribution of discoveries based on potential historical observations.

Popular topics associated with probability include randomness and uncertainty. It is asserted in Nassim Taleb's best-selling books The Black Swan and Fooled by Randomness that unusual occurrences often have greater significance than regular ones since their effect sizes are less constrained. Additionally, findings are unlikely to be determined because of their rarity.

Notes

Practical Uses for Probability Theory

Data scientists frequently use probability to simulate scenarios when trials carried out under comparable conditions produce varied results (such as when tossing a coin or a dice).

It is also very useful in the corporate sector. Consider the insurance sector, where actuarial records show the life expectancy of people a particular age. Instead, than making predictions about what will happen to any one person, the goal is to record an outcome that affects a large number of people collectively.

Similar approaches have been used in genetics, where predicting the likelihood of a hereditary disease based on assumptions about a particular person rather than the frequency of occurrence is tied to the disease's likelihood.

Testing the effectiveness of a new vaccination is one example, such the poliomyelitis testing done for the Salk vaccine in 1954 involving nearly two million kids. The vaccine, developed by the U.S. Public Health Service, virtually made polio a thing of the past in the developed world.

What Are the 3 Types of Probability?

Three categories of probability are frequently employed to collect information for statistical inference. These are:

1. Classical

   This style of probability is also referred to as the axiomatic technique and has a set of axioms (laws) associated with it. For instance, you might establish a rule that states that a probability must be higher than 0.5% to be valid.

2. Relative Frequency

   This entails examining the ratio of the frequency of a single event to all possible outcomes. After data from an experiment have been collected, this kind of probability is frequently used to compare a subset of data to the overall quantity of data obtained.

3. Subjective Probability

   Probability is the possibility of something occurring based on one's experiences or personal judgement when employing the subjective approach. Since subjective probability is dependent on one's opinions, judgements and personal reasoning, there are no formal calculations made here.

Probability Theory Examples

Researchers, corporations, financial analysts and countless others use probability theory as a tool for risk management and scenario analysis.

1. Epidemiology

   Consider epidemiology, the study of disease transmission. Researchers in this area examine the prevalence of diseases and how the likelihood varies among various population groupings. Today's use of probability by epidemiologists to evaluate the cause-and-effect link between exposure and disease from the coronavirus is an example of this.

   To identify important variables indicating the connection between exposures and health hazards, probability theory is frequently applied. Here, quantifying uncertainty

Notes

is the goal. This information can motivate a plan of action based on what will benefit persons suffering from various ailments the most.

2. Insurance

Actuaries, who are frequently employed in the insurance sector, use probability, statistics and other data science methods to determine the likelihood that a series of uncertain future occurrences will occur. The amount of money that needs to be set aside to cover potential losses is then calculated using various data concepts.

3. Small Business

Then there is the realm of small business, where proprietors are not always able to rely on their gut feelings and intuition to run a successful enterprise. Probability analysis can give business owners crucial indicators indicating the most lucrative and effective routes in today's cutthroat business market. This research provides a controlled method to predict future outcomes.

The graph will start with $500,000 at the low end and $750,000 at the high end, for instance, if a company anticipates monthly revenue of between $500,000 and $750,000. The graph will resemble a bell curve for a typical probability distribution, with the most likely outcomes being closer to the midpoint of the extremes and the least likely outcomes being closer to the extreme ends of the range.

4. Meteorology

For instance, if a company anticipates monthly sales of between $500,000 and $750,000, the graph will start with $500,000 at the low end and $750,000 at the high end. In a typical probability distribution, the graph will resemble a bell curve, with the most likely outcomes being closer to the middle of the extremes and the least likely outcomes being closer to the extreme ends of the range.

A numerical expression of uncertainty regarding the quantity or event being projected is included in a probability forecast. Ideally, a weather prediction would include information that precisely measures the inherent uncertainty in every component (temperature, wind, precipitation, etc.).

Surveys repeatedly show that customers want to know how confident or uncertain weather forecasts are. Because users can make decisions that explicitly account for this uncertainty, the widespread dissemination and good communication of prediction uncertainty information is likely to produce significant economic and societal advantages.

Advantages and Disadvantages of Probability Theory

For data scientists, there are several advantages and disadvantages with probability that need to be considered.

1. Classical

The traditional method of probability is used in the coin flip illustration above. For individuals without a background in maths or science, the classical approach provides an approach to real-world examples that is straightforward and simple to understand.

Regarding restrictions, the traditional method cannot manage tasks where there are an infinite number of potential outcomes. In situations where each outcome is not equally likely, such as when rolling a weighted dice, it is also ineffectual. These drawbacks limit this strategy's capacity to handle more challenging tasks.

2. Relative Frequency

Notes

Relative frequency has the benefit over the traditional technique in that it can manage situations where distinct events have varying theoretical probabilities (or likelihoods) of happening. This strategy can also handle a probability scenario when there are no known outcomes.

Relative frequency probability has some restrictions, even though it can be used in more contexts and circumstances than classical probability. Relative frequency is first constrained by the issue of "infinite repetitions." This is the point at which this theory cannot be used to analyse trials that have occurred an unlimited number of times. Even yet, there can only be a finite number of experiments undertaken.

3. Subjective

Problems that require some level of belief to be conceivable are those that benefit from subjective probability. For instance, a candidate who may be down in the polls could utilise subjective probability to support his or her decision to continue the campaign.

The so-called reference class problem is also advantageous for subjective probability. In a reference class problem, classifying an event could be necessary before assigning a probability to it. Since that categorisation may be arbitrary, modifying it may alter the likelihood that the event will occur.

For instance, identifying the classes of persons who are pertinent to the issue is the first step in calculating the likelihood that a person may develop an infectious disease like COVID-19. Here, different reference classes can be created.

The phrase "all U.S. residents" or a similar phrase could be used. Alternatively, the scope may be reduced to, say, "all residents of the states of X, Y and Z, where 80% of the deaths are occurring." In other words, various probabilities will emerge based on the reference class used.

3.3.2 Reproducibility

Measurements must be reproducible and consistent. It is the extent to which a tool can provide the same result when used again in identical circumstances. The terms repeatability and dependability are used interchangeably with the term reproducibility.

Even if a measurement has poor validity, it can nonetheless be highly reproducible. Contrary to popular belief, a measurement with good validity cannot also have strong reproducibility.

The terms replicability and reproducibility are used inconsistently and, in some circumstances, contradictorily by various scientific disciplines and institutions. Due to the lack of a consensus definition for this phrase, it becomes difficult to evaluate reproducibility.

However, achieving consistent results while employing the same input variables, methodological and computational methods and analysis settings can be characterised as reproducibility in the biological sciences.

Replicability, which is the process of achieving consistent results across research that addressed the same scientific inquiry and each of which has gathered its data, is closely related to the term reproducibility.

Reproducibility can be described using a multi-tiered method, according to the American Society for Cell Biology, which reflects how the term is understood among many scientific groups.

Notes

1. Direct replication: tries to duplicate an observed result by employing the same experimental setup and guidelines as the original study.
2. Analytic replication: recreating a collection of scientific discoveries by doing a new study on the initial data.
3. Systemic replication: recreating a scientific discovery under new experimental circumstances (for instance, utilising a new animal model or cell culture system)
4. Conceptual replication: employing a distinct set of experimental settings or techniques to assess the validity of the findings.

Measuring Precisely

It involves any kind of scientific observation, including measuring and counting. Scientific measurements are diverse and may include various kinds: Time, space dimensions, electric current, material qualities, acidity and concentration are among the parameters that can be measured.

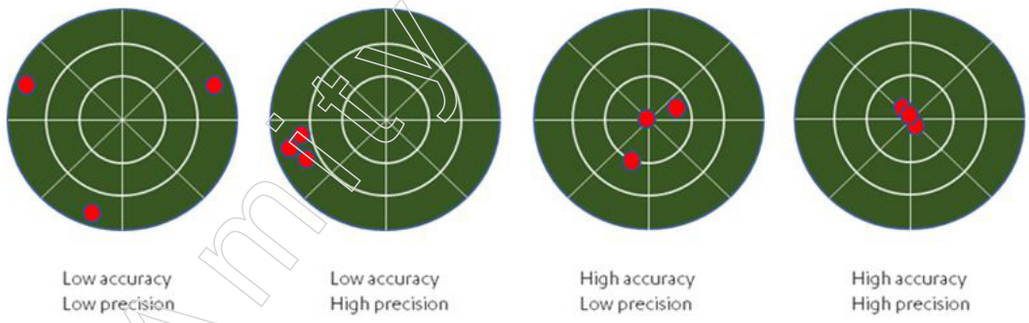
These are only a few instances from the natural sciences, not an exhaustive list. For instance, scientific observations in the social sciences are abundant in different counts and metrics across different fields.

Each measurement has a corresponding margin of error or uncertainty, which affects how definite it is. Uncertainties accompany all counts, measurements and other quantification methods; this is a fundamental aspect of scientific measurements.

The degree of measurement closeness that reduces the margin of error associated with the scientific observation is referred to as precision. The scale at which an observation is measured affects how close something is.

For example, the measurement of lengths spans numerous orders of diminishing magnitude, from metres to centimetres to millimetres, to microns, nanometres and angstroms. The exactness of something grows as it gets smaller and it's easier to pinpoint how close one measurement is to another.

Precision is also distinct from the accuracy of a measuring tool and can be illustrated as follows:



The first example shows low accuracy and precision since the three dots are on the outside ring, apart from one another and far from the bullseye. The second scenario shows excellent precision but low accuracy since the dots are all grouped together in a narrow band on the outer ring. The third example shows excellent accuracy but low precision since the dots are near to the bullseye but not close to one another.

The dots are finally near to the target of the bullseye in the fourth scenario. The result is quite precise and accurate in this final version. A data set's reproducibility can be determined by visually examining its means and standard deviations.

Notes

Reproducible Research

Open science's primary tenet is reproducible research. By using this process, a thorough explanation of the technique used to get the data is provided, making it simple to access and reproduce.

All data and files must be precisely labelled, sorted and documented for research efforts to be computationally replicable. But because methodologies are hard to come by, reproducibility in science remains a problem.

Factors Affecting Reproducibility in Life Science Research

A fundamental tenet of scientific inquiry across all types of fields is the independent verification of facts. Researchers must be able to replicate the results of published studies to self-correct and improve both the quality of the available data and the body of prior research.

The goal of reproducibility in any given field of study is to ensure the transparency of the procedures used to get at the results, not necessarily to guarantee the accuracy of the results themselves. In a perfect world, scientists would be able to repeat tests, produce the same outcomes and draw the same conclusions.

This isn't always the case; biological research, for instance, is impossible to duplicate due to the wide range of approaches used. This calls into question the validity of scientific findings and despite increased attention to the issue, research trainees and students continue to be unaware of the lack of repeatability.

The Reproducibility Problem

Over 70% of biologists were found to be unable to replicate the results of other scientists, according to a 2016 nature survey. Additionally, almost 60% of studies were unable to replicate their results.

This lack of reproducibility lowers the effectiveness of scientific output, impedes scientific advancement and increases the amount of time and money lost, all of which have a negative impact on the public's confidence in scientific research.

In fact, the cost of non-reproducible research, or the amount of money spent on preclinical research that cannot be replicated, is estimated to be $28 billion annually.

Overall, reproducibility is crucial to producing solid, amazing research that supports scientific advancements. Reproducibility barriers in the biological sciences have been shown to include several elements. Several guidelines and recommendations have therefore been developed, although it can be difficult to really put these ideas into practise.

The scientific community must adopt an impartial attitude to experiment design and be willing to represent and publish their results accurately and comprehensively, with precise descriptions of the procedures utilised, to assure reproducibility in the future.

Additionally, identifying the lack of reproducibility and advancing better research procedures and incentives in the life sciences are the responsibility of publishers, funders and politicians. In the end, this can aid in enhancing research procedures and ensuring the strong credibility of scientific findings.

3.3.3 String Processing

Data wrangling, which is changing and modifying data to make it acceptable for

Notes

analysis, heavily relies on string processing. Here are a few typical string processing methods for data wrangling:

1. Cleaning and formatting:

   String data frequently has flaws and inconsistencies such leading or trailing spaces, special characters, or erroneous capitalization. Data quality is increased and consistency is ensured by using cleaning and formatting procedures such eliminating punctuation, changing to lowercase or uppercase and standardising formats.

2. Tokenization:

   Tokenization is the process of dividing a string into tokens, which are typically words or sentences. When you need to analyse text in detail, this method is helpful. Tokenizing a text, for instance, enables you to count word frequencies or examine sentiment.

3. Splitting and joining:

   A string can be divided into substrings or combined from multiple strings using the splitting and joining procedures. For example, you can break comma-separated values from a string into separate values or combine numerous columns into one column.

4. Regular expressions:

   Regular expressions (regex) offer a potent and adaptable method for finding and modifying strings based on patterns. Regex can be used to substitute text, validate string formats and identify and extract specific patterns. For instance, using regex to extract email addresses or phone numbers from a text.

5. String matching and searching:

   Searching for certain patterns or substrings inside a longer string is known as string matching. Common matching methods include precise matching, partial matching and fuzzy matching. With the aid of these methods, pertinent data can be found and extracted from unstructured or semi-structured text data.

6. String manipulation:

   Programming languages and libraries offer a variety of string manipulation capabilities and techniques. Concatenation, substring extraction, string substitution, padding and truncation are a few examples of these. With these procedures, you can restructure and alter strings to suit your data needs.

7. Encoding and decoding:

   Transforming strings from one character encoding system to another is known as encoding. In addition to ASCII, Unicode and UTF-8, there are several common encoding methods. Decoding is the opposite operation. When working with data in many languages or character sets, these methods are crucial.

8. Text normalization:

   Text normalisation entails putting text into a consistent format. It incorporates methods like lemmatization (reducing words to their base or dictionary form), stemming (reducing words to their root form) and deleting stop words (common words with minimal semantic significance). Textual data's dimensionality and noise can be reduced with the aid of text normalisation.

9. String comparison:

   Data wrangling operations frequently call for string comparisons. String comparison can involve comparing strings for equality, alphabetizing strings, or figuring out how

Notes

similar two strings are (using techniques like edit distance or cosine similarity, for example). Tasks involving deduplication, record linking, or grouping can benefit from these comparisons.

These are but a handful of instances of string processing methods used in data wrangling. The techniques chosen to rely on the precise data and analysis objectives. A wide range of tools are available in programming languages and libraries, such as Python's built-in string methods, regular expression libraries (such as re), or text processing libraries (such as NLTK, spaCy), to carry out these tasks effectively.

3.3.4 Dates, Times and Text Mining

Dates are represented by the Date class in R, while Times are represented by the POSIXct or POSIXlt class. The as. In R, dates are handled using the Date() function. The date is entered using the format YYYY-MM-DD or YYY/MM/DD as a String and the internal representation of the date by this function is the number of days since January 1, 1970. Additionally, internal timekeeping records the seconds since January 1, 1970.

Facts of Dates and Times in R:

1. Dates are represented by the Date class.
2. Times are represented by the POSIXct or the POSIXlt class.
3. Dates are stored internally as an integer representing the number of days since 1970-01-01.
4. Times are stored internally as an integer representing the number of seconds since 1970-01-01.

Text Mining

Text mining is one of the most crucial techniques for analysing and processing unstructured data, which accounts for over 80% of all data worldwide. In data warehouses and cloud platforms, which are used by many businesses and institutions today, vast volumes of data are collected and stored. As new data is constantly being added from new sources, this data is growing drastically by the minute.

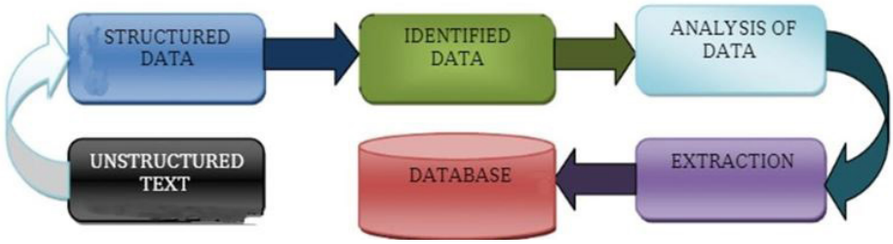
Large amounts of textual data are therefore difficult for businesses and organisations to store, manage and analyse using conventional technologies. Developing your data science skills will enable you to overcome the difficulties. Let's discuss text mining in greater detail.

The five fundamental steps involved in text mining are:

1. Collecting unstructured data from many data sources, including blogs, emails, plain text, web pages, PDF files and other sources.
2. Perform pre-processing and purification procedures on the data to find and eliminate anomalies. Data cleansing enables you to retrieve and preserve the priceless information concealed within the data as well as to determine the linguistic roots of phrases.
3. Several text mining tools and applications are provided for this.
4. Transform all the pertinent data that was gleaned from unstructured data into formats that are structured.

Notes

5. Use the Management Information System (MIS) to analyse the data trends.
6. Put all the important data in a safe database to facilitate trend analysis and improve the organisation's decision-making process.



Text Mining Techniques

The procedures used to glean insights from text can help one better understand text mining approaches. A range of text mining tools and software are commonly employed for the implementation of various text mining methodologies. Let's look at the different text mining techniques now:

Now let's examine the most well-known text mining approaches:

1. Information Extraction

   The most popular text mining method is this one. Extraction of pertinent information from vast quantities of text is the process of information exchange. Using semi-structured or unstructured texts, this text mining technique aims to extract entities, properties and their relationships.

2. Information Retrieval

   Information retrieval (IR) is the process of gathering pertinent and related patterns from a specific set of words or phrases. IR systems use several algorithms in this text mining technology to detect and analyse user actions and, as a result, find pertinent data. Google and Yahoo are the most well-known IR systems. The most well-known text mining method is information retrieval.

3. Categorization

   With this method, documents written in natural language are categorised according to their content and put into one of a predetermined number of subject categories. It is one of the "supervised" learning techniques used in text mining. Categorization, or more specifically Natural Language Processing (NLP), is the method of gathering the content of the documents and processing and analysing them to determine the suitable subjects or indexes for each text document.

4. Clustering

   One of the most important text mining approaches is clustering. It looks for inherent textual informational structures and classifies them into useful subgroups or 'clusters' for additional examination. Creating meaningful clusters from unlabelled textual data without any prior knowledge of them is a difficult task in the clustering process.

5. Summarisation

   Text summarising is the discipline of automatically producing a compacted version of a certain text that includes information that is helpful to the end user. The objective of this text mining technique is to look through a variety of text sources to find texts that contain a lot of information and summarise them in a concise manner while mainly keeping the overall substance and aim of the original documents.

Notes

Technique	Characteristics	Tools
Retrieval	Retrievals valuable information from unstructured text	Intelligent Miner, Text Analyst
Extraction	Extract information from structured database	Text Finder, Clear Forest Text
Summarization	Reduce length by keeping its main points and overall meaning as it is	Tropic Tracking Tool, Sentence Ext Tool
Categorization	Document based categorization	Intelligent Miner
Cluster	Cluster collection of documents, Clustering, classification and analysis of text document	Carrot, Rapid Miner

3.4 Data Bias

Data bias in finance refers to systematic errors or prejudices in financial datasets, which can lead to unfair or inaccurate outcomes. It can arise from various sources, such as historical, incomplete data collection, or algorithmic biases.

Financial data bias can have serious repercussions, such as distorted investment decisions, unfair pricing, or biased lending decisions. To achieve fair and equitable outcomes in financial systems, it is crucial to recognise and mitigate data bias. This calls for rigorous data collection, preprocessing and algorithmic design.

Machine learning is increasingly concerned with data bias, which can take many different forms and manifest itself in the collection, analysis and use of data. Biased data produces unreliable and erroneous outcomes, making it ineffectual at attaining intended goals and possibly harming people.

3.4.1 Bias in Data Science

Decisions are increasingly being made using or informing by machine learning algorithms. A model might, for instance, have an impact on a decision about the granting of a loan or the screening of applicant resumes for a job application. These choices are critical; therefore must be sure that our models don't make any distinctions based on things like race, gender, or age.

Unintentional bias can frequently be found in machine learning models, which can lead to unreliable and unfair conclusions. A solid machine learning model involves more than merely computing loss measures to build and evaluate. It is crucial to examine your training data and occasionally the source of the data to check for biases before operationalizing a model.

1. Reporting Bias:

Selective reporting, often referred to as reporting bias, occurs when only a subset of results or outcomes are included in a data set, which typically represents a small portion of the total amount of real-world data. People have a propensity to report less information than is accessible.

Notes

Types of reporting bias -

- Citation bias: when your analysis is founded on studies that are cited in other papers.
- Language bias: occurs when you disregard news articles that are not in your native tongue.
- Duplicate publication bias: occurs when studies that are published in several places are given higher weight.
- Location bias: arises when some research are more difficult to find than others.
- Publication bias: happens when research with favourable findings is more likely to be published than research with unfavourable or insignificant findings.
- Outcome reporting bias: occurs when specific outcomes are only sometimes reported. For instance, in a quarterly report, you only report when the company produces positive results.
- Time lag bias: when studies take years to publish, this happens.

2. Automation Bias

Automated bias is the propensity for people to favour outcomes or recommendations produced by automated systems and to disregard contrary data from non-automated systems, even when it is accurate.

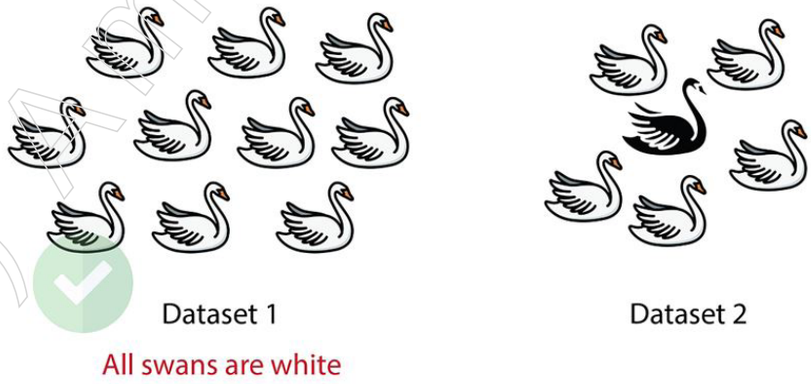
3. Selection Bias

Selection bias arises when the method that data are chosen is not accurate to the distribution of data. This occurs because of improper randomization during data collection.

Types of selection bias -

- Sampling bias: occurs when randomization while data collection is improperly achieved.
- Convergence bias: when information is not chosen in a representative way. For instance, if you survey only the half of your consumers who have purchased your product and not the other half, your dataset will not accurately reflect the group of people who have not made a purchase.
- Participation bias: when participation gaps in the data collection process cause the data to be unrepresentative.

4. Overgeneralization Bias



Notes

Regardless of the size of the dataset, overgeneralization happens when you think that what you observe in your dataset is what you would observe in any other dataset intended to evaluate the same information.

5. Group Attribution Bias

People frequently generalise an entire group based solely on the actions of a small number of its members. Group attribution bias is the phrase used to describe this propensity to generalise the truths about people to the entire group to which they belong.

Types of Group Attribution Bias -

- In-group bias: occurs when you offer priority to those who are a part of a group that you are a personal member of or who share interests with. For instance, a manager writing a job description for a data scientist post might assume that qualified candidates must possess a master's degree since he or she does (regardless of their work experience).
- Out-group bias: if you stereotype certain people of a group to which you do not directly belong. For instance, a manager with a master's degree who is writing the job description for a data scientist post thinks that candidates without a master's degree lack the necessary knowledge for the position.

6. Implicit Bias

When assumptions are drawn from personal experiences that do not necessarily apply to more generic situations, implicit bias results. People frequently act in ways that are biassed and stereotypical without realising it.

For e.g., Red is flagged as dangerous by a North American computer vision engineer. Red, however, is a well-liked colour in Chinese culture and represents luck, joy and happiness.

3.4.2 Overcoming Bias

Machine learning (ML) and artificial intelligence (AI) have quickly progressed from novel notions to indispensable tools for brand marketers. According to Salesforce, 68 out of marketers claim to have a fully developed AI strategy, an increase from 57% in 2020 and 60% in 2021. It is essential to predictive modelling, campaign personalisation, CX optimisation and almost all other marketing-related processes.

But as many businesses have learned over the years, AI has its drawbacks. A software developer found that Apple Card's algorithm was inherently sexist in 2019. The fintech sector received criticism for discriminating against people of colour in mortgage loans and home refinancing during the same year. Despite the apparent differences between the two occurrences, data bias is their common fundamental cause.

The Problem with Biased Data

Although biased data is harmful in and of itself, the consequences that follow are much worse. As they say, "Garbage in, garbage out." According to study respondents for Forrester Consulting, inaccurate data caused them to squander more than 20% of their marketing spend. There are certain specific ways data bias may cause havoc for brand marketers:

Missed opportunities: You cannot afford to make decisions based on skewed data in the digital world since it moves at the speed of light. Inadequate insights might cause you and your team to miss out on conversion, upsell and retention chances.

Notes

Skewed customer journey insights: To enhance clients' experiences with your brand, it is essential to comprehend their travels. Examining the various touchpoints that make up a customer journey is complicated enough, but if you distort the process with inaccurate data, you won't be able to properly meet the demands of your users.

Ineffective marketing campaigns: For marketers, boosting ROI is a constant issue. As you optimise and adjust based on inaccurate insights and false assumptions, biased data will only make that situation worse.

Compliance violations: AI can make the already difficult task of complying with GDPR and other privacy laws much more difficult. Consumers are given special rights under the GDPR regarding the use of their data for automated decision-making or profiling. You run the risk of facing hefty fines if your automated systems improperly profile customers based on biased data.

Exacerbated organisational bottlenecks: Too frequently, IT or data scientists are forced to submit information to marketing departments. Data bias causes additional work for data scientists to clean up, which causes additional delays for you to receive the data you need.

Why debiasing data matters

Data debiasing can be a difficult, unreliable procedure. Even so, you could be considering giving up. Although it requires a lot of work, the consequences of putting it off can be much worse. Data bias can—and will—damage your brand, whether it's through high fines for privacy violations or declining ROI.

1. Integrate all your brand's data so that you can use it without overlooking anything crucial.
2. End-to-end privacy controls guarantee the security and safety of your users' data.
3. Interactive, real-time dashboards offer ongoing insight into the paths taken by your customers.
4. Resources for executing sophisticated queries with minimal technical expertise.
5. Democratise data across the board so that departments can concentrate on what they do best.
6. Easily maintain client data flow into the system to guarantee AI algorithms always have access to up-to-date, accurate data.

3.4.3 Bias Alerts

Bias alert is a proactive approach that aims to identify and address biases throughout the data science lifecycle. It involves developing techniques, tools and frameworks to detect, measure and mitigate biases, ensuring fair and ethical decision-making. The key objectives of bias alert are:

1. Awareness and Understanding: Bias alert raises decision-makers' and data scientists' awareness of potential biases in data science procedures. It promotes a thorough comprehension of the moral and societal repercussions of biased algorithms.
2. Bias Detection: Bias alert systems must include techniques and methods for detecting bias. These methods examine models and data to find trends, correlations and possible biases. They can alert data scientists to potential biases by highlighting those instances.

Notes

3. Bias Mitigation: Beyond only identifying biases, bias alert works to lessen them in data science procedures. You can use approaches like algorithmic fairness, debiasing techniques and counterfactual fairness to minimise or get rid of biased results. More inclusive and objective decision-making can also be achieved by utilising a variety of data sources, varied teams and user feedback.

Implementation Strategies

To incorporate bias alert effectively, organisations and data science practitioners can adopt the following strategies:

1. Data Governance
2. Model Evaluation
3. Continuous Monitoring
4. Ethical Considerations

Biases must be minimised or eradicated as data science continues to impact our society to promote justice, equality and moral decision-making. With its proactive approach to identifying, measuring and mitigating biases at every stage of the data science lifecycle, bias alert plays a crucial part in this quest. Organisations may help create more responsible and inclusive data science systems by deploying bias alert mechanisms, working towards a just and equitable future.

3.5 Machine Learning Algorithms

IIoT machine learning algorithms can lead to financial savings, quicker processing and improved performance. We've recently seen the benefits of machine learning methods from streaming movie services that recommend films based on viewing preferences to monitor fraud based on client spending trends.

To find or learn about underlying patterns in data, machine learning algorithms are techniques for mathematical model mapping. Machine learning is a class of computational algorithms that can identify patterns in data, classify it and make predictions by learning from previously collected data (training set).

3.5.1 Machine Learning Algorithms

Machine learning algorithms can be used in programmes to find hidden patterns in data, forecast outcomes and improve performance based on prior performance. For a variety of tasks, many machine learning algorithms can be used, such as basic linear regression for prediction problems like stock market forecasting and the KNN algorithm for classification problems.

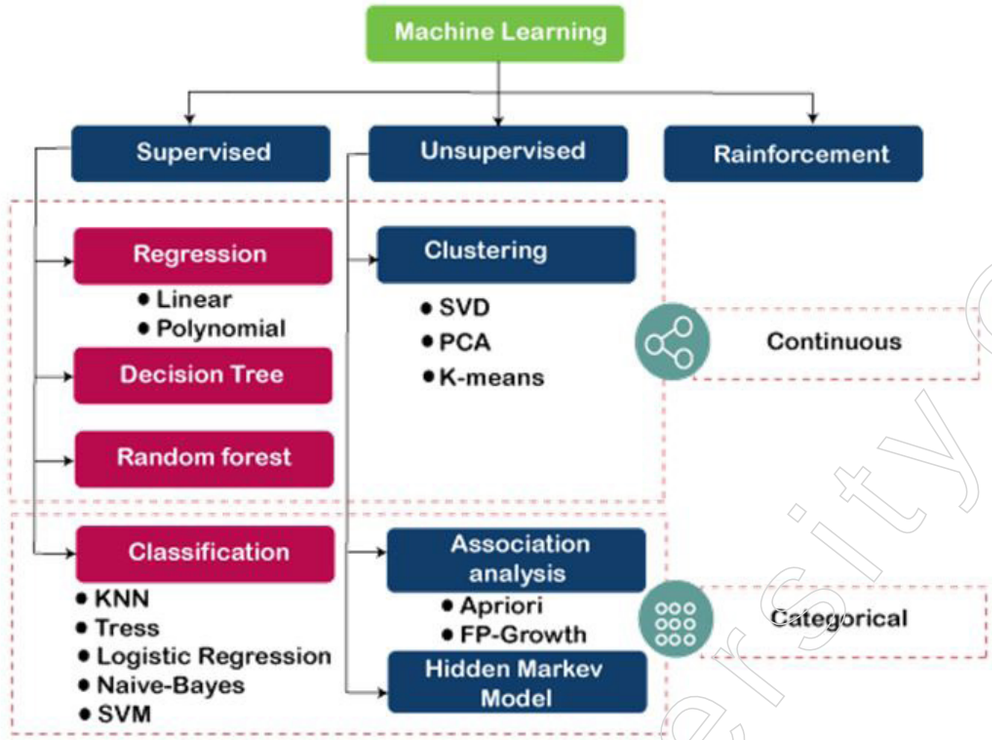
Types of Machine Learning Algorithms

Machine Learning Algorithm can be broadly classified into three types:

1. Supervised Learning Algorithms
2. Unsupervised Learning Algorithms
3. Reinforcement Learning algorithm

The below diagram illustrates the different ML algorithm, along with the categories:

Notes



1) Supervised Learning Algorithm

A computer needs external supervision to learn through supervised learning. To train the supervised learning models, the labelled dataset is used. The model is assessed by sending a sample of test data after training and processing is complete to see if it predicts the desired outcome.

The basic goal of supervised learning is to map input and output data. Because supervised learning depends on supervision, it is the same as when a student studies under a teacher's guidance. One example of supervised learning is spam filtering.

Supervised learning can be divided further into two categories of problem:

- Classification- In order to accurately classify test data into different categories, classification uses an algorithm. It identifies particular entities in the dataset and makes an effort to determine how those things should be defined or labelled. Linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbour and random forests are typical classification algorithms.
- Regression- To comprehend the relationship between dependent and independent variables, regression is used. It is frequently used to produce estimates, including those for a company's sales revenue. Popular regression algorithms include linear regression, logistical regression and polynomial regression.

Simple linear regression, decision trees, logistic regression, the KNN method and others are prominent supervised learning algorithms.

2) Unsupervised Learning Algorithm

Unsupervised learning is a type of machine learning where the computer can come to its own conclusions based only on the information that is provided. Unsupervised

Notes

models can be trained using the unlabelled dataset, which is neither categorised nor classed and the algorithm must work on the data without any supervision.

In unsupervised learning, instead of providing a planned outcome, the model sifts through the enormous amount of data in quest of insightful understandings. To deal with the Association and Clustering problems, they are used. As a result, it can be split into two groups:

- ❖ Clustering
- ❖ Association

K-means Clustering, Apriori Algorithm, Eclat and other algorithms are examples of certain unsupervised learning methods.

3) Reinforcement Learning

When using reinforcement learning, an agent creates interactions with its surroundings and gains knowledge from the responses. The agent is given feedback in the form of rewards; for instance, he is given a positive reward for every successful activity and a negative reward for every unsuccessful activity. The agent is not overseen in any way. The Q-Learning algorithm is used in reinforcement learning.

List of Popular Machine Learning Algorithm

1. Linear Regression Algorithm
2. Logistic Regression Algorithm
3. Decision Tree
4. SVM
5. Naïve Bayes
6. KNN
7. K-Means Clustering
8. Random Forest
9. Apriori

3.5.2 Linear Regression

Another form of machine learning technique is linear regression, specifically supervised machine learning, which learns from labelled datasets and converts data points into the best-fitting linear functions. This is useful for making predictions using new datasets. First need to be familiar with supervised machine learning algorithms. The algorithm learns in this type of machine learning from labelled data. Data that has been labelled is a dataset whose corresponding target value is already known. Supervised learning has two types:

- Classification: Based on the independent input variable, it forecasts the dataset's class. Categorical or discrete values are known as classes. like the representation of an animal as a dog or cat.
- Regression: Based on the independent input variable, it forecasts the continuous output variables. like the estimation of property prices depending on various factors, such as the age of the house, its distance from the main road, its location, its neighbourhood, etc.

Notes

The linear relationship between a dependent variable and one or more independent features is determined using supervised machine learning's "linear regression" method. Univariate linear regression occurs when the number of independent features is 1, but multivariate linear regression occurs when there are several independent features.

Finance, economics and psychology are just a few of the disciplines that employ linear regression to analyse and forecast the behaviour of a given variable. For instance, in the field of finance, linear regression may be used to comprehend the connection between a company's stock price and earnings or to forecast the value of a currency based on its historical performance.

Regression is one of the most significant supervised learning tasks. In regression, a series of records with X and Y values are present and these values are used to train a function that may be used to predict Y from an unknown X. Regression requires us to determine the value of Y, hence needed a function that can forecast continuous Y when X are independent features.

Assumption for Linear Regression Model

However, for it to be accurate and dependable, linear regression must meet a few requirements. Linear regression is a strong tool for understanding and forecasting the behaviour of a variable.

1. Linearity: The relationship between the independent and dependent variables is linear. This suggests a linear relationship between changes in the dependent variable and those in the independent variable or variables.
2. Independence: The dataset's observations are not dependent on one another. The dependent variable's value for one observation does not depend on the dependent variable's value for another observation, according to this.
3. Homoscedasticity: The variance of the mistakes remains constant for all independent variable (or independent variable(s)) levels. This demonstrates that the size of the independent variable(s) has no bearing on the errors' variance.
4. Normality: The model's errors are evenly distributed.
5. No multicollinearity: The independent variables don't have a strong correlation with one another. This suggests that the independent variables have little to no association with one another.

3.5.3 Logistics Regression

Logistic regression falls under the genre of Supervised Learning and is one of the Machine Learning algorithms that is most frequently employed. It is employed to forecast the categorical dependent variable using a specific set of independent factors.

Apart from how they are applied, logistic regression and linear regression are very similar. difficulties involving regression are solved using linear regression, while classification difficulties are solved using logistic regression.

Logistic regression can be used to quickly find the factors that will function well when categorising observations using different sources of data.

Logistic Function (Sigmoid Function):

- A mathematical formula called the sigmoid function is used to convert expected values into probabilities.

Notes

- It transforms any real value between 0 and 1 into another value.
- The logistic regression's value must fall within the range of 0 and 1 and it is not allowed to exceed this number. As a result, it takes the shape of a "S" curve. The logistic function or Sigmoid function are other names for the S-form curve.
- In logistic regression, the threshold value concept is used to specify the probability of either 0 or 1. For example, values that are above the threshold value tend to be 1 and values that are below the threshold values likely to be 0.

Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

Logistic Regression Equation:

The mathematical steps to get Logistic Regression equations are given below:

Now aware that the straight line's equation can be expressed as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}; \text{ 0 for y= 0, and infinity for y=1}$$

But need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Type of Logistic Regression:

Three forms of logistic regression can be distinguished based on the categories:

1. Binomial: In binomial logistic regression, the dependent variables can only take one of two possible forms, such as 0 or 1, Pass or Fail, etc.
2. Multinomial: One of three or more possible unordered sorts, such as "cat," "dogs," or "sheep," may be the dependent variable in multinomial logistic regression.
3. Ordinal: Ordinal logistic regression allows for up to three different ordered types of dependent variables, such as "low," "Medium," or "High."

3.5.4 Dirty Data and Naïve Bayes

A supervised learning technique for classification problems, the Naïve Bayes algorithm is based on the Bayes theorem. It comes with a substantial training set and is primarily used for text classification.

Why is it called Naïve Bayes?

The words Naïve and Bayes, which are combined to form the Naïve Bayes algorithm, are:

Naïve: It is referred regarded as naive since it assumes that the existence of one

Notes

trait is unconnected to the prevalence of other features. For instance, if the fruit's red, spherical and sweet fruit, form and flavour identify it as an apple. Each attribute therefore functions independently of the others to aid individuals in identifying an apple.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem:

The Bayes theorem, commonly referred to as Bayes' Rule or Bayes' law, is used to calculate the likelihood of a hypothesis given some prior information. The conditional probability determines this.

The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Working of Naïve Bayes' Classifier:

Working of Naïve Bayes' Classifier can be understood with the help of the below example:

Assume that "Play" is the equivalent target variable and that have a dataset of weather conditions. Therefore, utilising this dataset, must determine whether to play on a specific day based on the weather. Thus, must take the following actions to remedy this issue:

- ❖ Create frequency tables from the provided dataset.
- ❖ Create a table of likelihoods by calculating the probability of specific attributes.
- ❖ To determine the posterior probability, utilise the Bayes theorem at this time.

Advantages of Naïve Bayes Classifier:

1. To forecast a class of datasets, Nave Bayes is one of the quick and simple machine learning methods.
2. It is applicable to both binary and multi-class classifications.
3. It outperforms the other algorithms in multi-class predictions.
4. It is the most often used solution for text classification issues.

Disadvantages of Naïve Bayes Classifier:

Assuming that all features are independent or unconnected prevents Naïve Bayes from learning the relationship between them.

Applications of Naïve Bayes Classifier:

- It's employed in credit scoring.
- It is utilised to classify medical data.

Notes

- Since Naïve Bayes Classifier is a quick learner, it can be utilised to make predictions in real time.
- Text categorization applications like spam filtering and sentiment analysis use it.

Types of Naïve Bayes Model:

There are three types of Naive Bayes Model, which are given below:

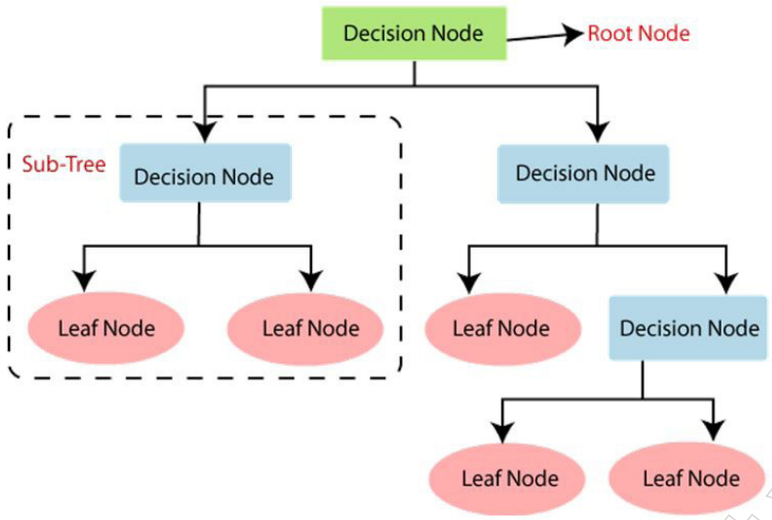
1. Gaussian: Characteristics are assumed to be distributed regularly in the Gaussian model. Accordingly, the model assumes that predictors use continuous values rather than discrete values and that these values are samples from the Gaussian distribution.
2. Multinomial: The Multinomial Naive Bayes classifier is used with multinomially distributed data. It identifies the categorization that a certain document fits within, such as Sports, Politics, Education, etc. and is frequently used to resolve problems with document classification. The classifier uses the predictor word frequency.
3. Bernoulli: Like a multinomial classifier, the Bernoulli classifier uses independent Boolean variables as predictor variables. such as determining whether a word is used or not in a document. This paradigm is well known for jobs involving document classification.

3.5.5 Decision Tree

• Decision Trees are a supervised learning technique that may be used to solve classification and regression problems, while it is often selected to solve classification problems. It is a tree-structured classifier, where internal nodes stand in for the dataset's characteristics, branches for the steps taken to arrive at conclusions and each leaf node for the classification result.

- The Decision Node and the Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and have no further branches, Decision nodes are used to create decisions and have multiple branches.
- Based on characteristics of the provided dataset, choices or tests are carried out.
- Depending on specified criteria, it offers a graphic representation of every option for a problem or decision.
- It has the shape of a tree-like structure because, like a tree, it begins at the root node and grows by adding more branches. This is why it is termed a decision tree.
- To build a tree, the Classification and Regression Tree algorithm, or CART algorithm, is utilised.
- A decision tree only asks a question and then divides into subtrees depending on the answer (Yes/No).
- Below diagram explains the general structure of a decision tree:

Notes



Why use Decision Trees?

Choosing the best approach for the dataset and task at hand is the most crucial consideration when creating a machine learning model. Below are the two advantages of using a decision tree:

- Since decision trees often mirror how people think while making decisions, they are easy to understand.
- The logic underpinning the decision tree is easy to understand because it exhibits a tree-like structure.

Decision Tree Terminologies

- Root Node: At the root node, the decision tree is initiated. The entire dataset has been divided into two or more homogeneous sets.
- Leaf Node: After obtaining a leaf node, the tree cannot be further divided because leaf nodes are the last output nodes.
- Splitting: The method of splitting involves separating the decision node/root node into sub-nodes in accordance with the specified conditions.
- Branch/Sub Tree: a tree created by slicing the original tree.
- Pruning: Pruning is the procedure of removing the tree's undesirable branches.
- Parent/Child node: The parent node of the tree and the remaining nodes are referred to as the child nodes.

3.5.6 K-Nearest neighbors (k-NN)

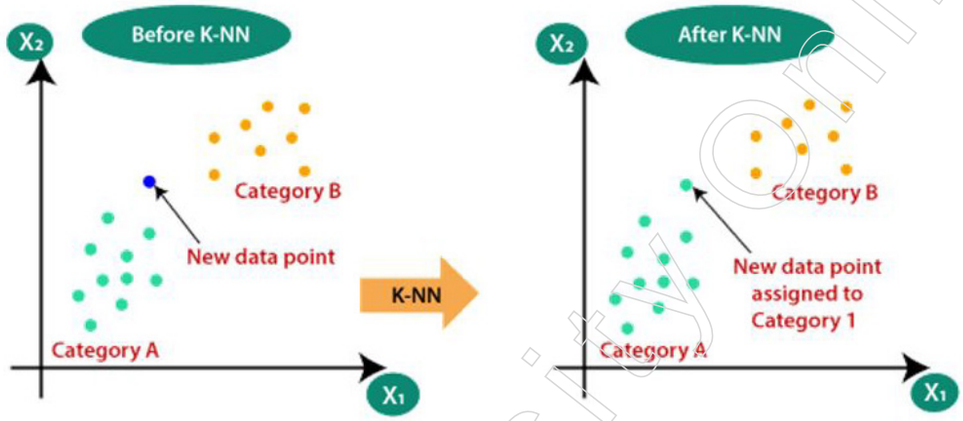
K-Nearest Neighbour is one of the most basic machine learning algorithms that uses the supervised learning approach. The K-NN technique places the new instance in the category that is most similar to the current categories, presuming that the new example and the previous cases are comparable.

Since it saves the training dataset instead of instantly learning from it, the algorithm is sometimes known as a lazy learner. Instead, it performs an operation on the dataset in order to classify the data. When classifying new data into a category that is roughly similar to the new data, the KNN algorithm simply saves the training dataset.

Notes

Why do a K-NN Algorithm is needed?

Which category does a new data item, x1, that is present in two categories, Category A and Category B, belong in? To handle this kind of problem, you need a K-NN algorithm. K-NN makes it simple to determine the category or class of a given dataset. Consider the below diagram:



Advantages of KNN Algorithm:

- Simple to implement.
- It can withstand noisy training data.
- If there is a lot of training data, it might be more effective.

Disadvantages of KNN Algorithm:

- K's value must always be determined, which might be difficult at times.
- The cost of calculation is considerable since it must calculate the separation between each data point for each training sample.

3.5.7 K- Means

K-Means Clustering, an unsupervised learning approach, is used to tackle the clustering problems in machine learning or data science.

The unlabelled dataset is split into k separate clusters using an iterative process and each dataset is only a member of one group that shares characteristics with the others.

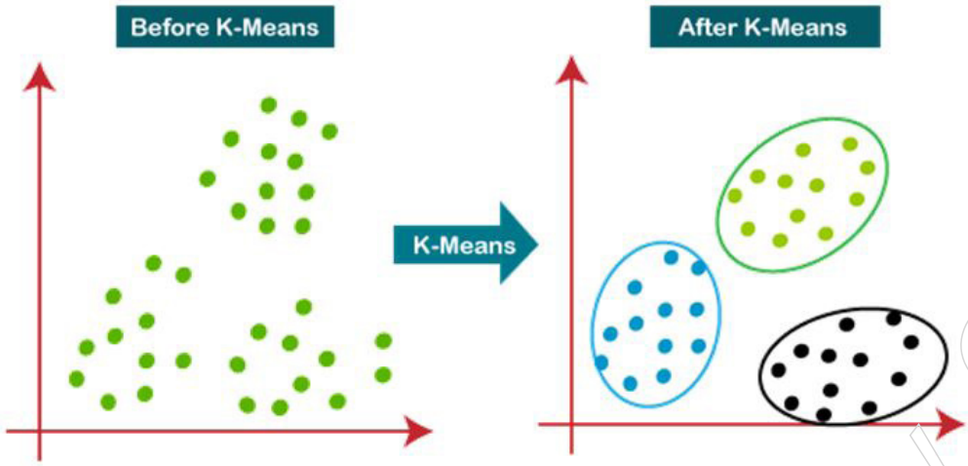
In general, the k-means clustering algorithm completes two jobs:

- Employs an iterative technique to find the optimal value for K centroids or centre points.
- Each data point is assigned to the nearest k-centre. A cluster is formed by the data points that are close to the k-centre.

As a result, each cluster stands out from the others and has some shared datapoints.

The K-means Clustering Algorithm is explained in the diagram below:

Notes

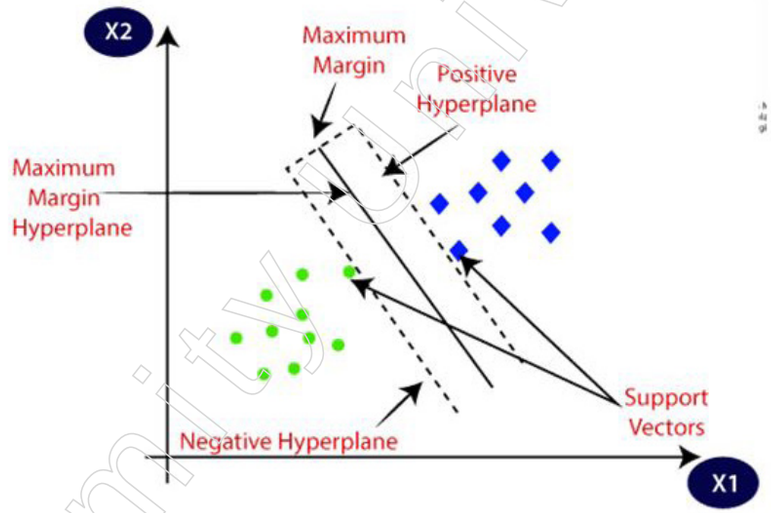


3.5.8 Support Vector Machine

One of the most popular supervised learning algorithms is called the Support Vector Machine, or SVM and it is used to solve Classification and Regression problems. However, it is primarily utilised in Machine Learning Classification issues.

With the help of the SVM algorithm, we can quickly categorise additional data points in the future by determining the best decision boundary or line that can divide up n-dimensional space. A hyperplane is the name given to this ideal decision boundary.

Look at the diagram below, where a decision boundary or hyperplane is used to categorise two distinct categories:



SVM algorithm can be used for Face detection, image classification, text categorization, etc.

Types of SVM

SVM can be of two types:

- Linear SVM: Data that can be linearly split into two groups or that can be separated into two groups using just one straight line are used with linear SVM. A Linear SVM classifier is used to categorise this data.
- Non-linear SVM: A dataset is said to have been "non-linearly separated" when a

Notes

straight line cannot be used to divide it into segments for classification and the used classifier is known as a non-linear SVM classifier.

3.5.9 Apriori

The Apriori algorithm is designed to work with databases that have transactional data and create association rules from frequent item sets. Using these association rules, it determines how strongly or weakly two objects are related. With the help of a hash tree and a breadth-first search, this approach quickly determines the relationships between the itemset. Iterative techniques are needed to identify the common item sets from a large dataset.

R. Agrawal and Srikant first presented this technique in 1994. It is mostly used for market basket analysis and aids in identifying products that can be purchased in combination. It can also be used in the healthcare sector to assist patients in identifying drug reactions.

Steps for Apriori Algorithm

Below are the steps for the apriori algorithm:

Step-1: Determine the support of itemset in the transactional database and select the minimum support and confidence.

Step-2: Take all supports in the transaction with higher support value than the minimum or selected support value.

Step-3: Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.

Step-4: Sort the rules as the decreasing order of lift.

3.6 Case study

Hotel Recommendation System

Typically, a hotel recommendation system uses collaborative filtering to provide recommendations based on feedback from users who fall within the same category as the user looking for a product.

Use Case: Everybody plans vacations and the initial step in any trip preparation is to locate a hotel. Numerous websites offer advice on which hotel would be best for our trip. A hotel suggestion system seeks to foretell which hotel among all hotels a user will most likely select.

Therefore, to create this kind of system that will assist the user in choosing the best hotel out of all the hotels. Customer testimonials will help us with this.

For example, Assuming you're planning a business trip, the hotel suggestion system should present the lodging options that previous guests have rated as the best for such trips. Building a recommendation system based on consumer feedback and ratings is thus also our strategy. So create a hotel suggestion system using the ratings and reviews provided by users who fall into the same category as the user.

Notes

```
recommend_hotel('UK','I am going on a honeymoon, I need a honeymoon suite room for 3 nights
```

	Hotel_Name	Average_Score	Hotel_Address
0	Haymarket Hotel	9.6	1 Suffolk Place Westminster Borough London SW1...
1	41	9.6	41 Buckingham Palace Road Westminster Borough ...
2	Taj 51 Buckingham Gate Suites and Residences	9.5	Buckingham Gate Westminster Borough London SW1...
3	Charlotte Street Hotel	9.5	15 17 Charlotte Street Hotel Westminster Borou...
4	Ham Yard Hotel	9.5	One Ham Yard Westminster Borough London W1D 7D...

The output of the Hotel Recommendation System

Summary

- Data munging and data cleanup are some names for it. Prior to performing any data analysis, you should often go through the data wrangling procedure to make sure the data are accurate and comprehensive
- Before data is ready for analytics, data wrangling software typically goes through six iterative steps: discovering, structuring, cleaning, enriching, validating and publishing.
- A critical phase in the data analytics process is data wrangling, which involves converting raw data into a more useful and intelligible form for additional analysis.
- Web scrapers, a type of scraping software, are used to carry out web scraping.
- Web scraping is the main cause of the rise in copyright violations, terms of service violations and other activities that are seriously detrimental to a company's operations, even while it technically speeds up data surfing, loading, copying and pasting.
- Data-scraping in combination with big data, can give the organisation market information, assist them uncover important trends and patterns and identify the greatest prospects and solutions because we are on the cusp of a data transformation.
- Dirty data, in essence, is erroneous information that interferes with a company's database and has an impact on critical operations like GTM, segmentation, customization, lead scoring, prospecting and planning for optimal customer profiles, among others.
- Probability is a key subject that new data scientists must understand. Numerous crucial data science concepts, from inferential statistics to Bayesian networks, are built on notions from probability theory.
- Data scientists can evaluate the likelihood of results of a specific study or experiment using probability. An experiment is a planned investigation that is carried out under supervised circumstances. The experiment is referred to as a chance experiment if the outcome is unpredicted.
- Replicability, which is the process of achieving consistent results across research that addressed the same scientific inquiry and each of which has gathered its data, is closely related to the term reproducibility.
- One of the most important methods for assessing and processing unstructured data, which makes up almost 80% of all data in the world, is text mining. Many

Notes

organisations and institutions today collect and store enormous volumes of data in data warehouses and cloud platforms and as fresh data floods in from various sources, this data continues to expand dramatically by the minute.

- The Naïve Bayes algorithm is comprised of two words Naïve and Bayes.

Glossary

- Data wrangling: The process of transforming unusable data into a useful form is known as data wrangling.
- Web scaping: The process of gathering data from webpages is called web scraping
- Parsing: Making anything comprehensible for component-by-component analysis is referred to as parsing.
- Dirty Data: The one bad fruit that can ruin your entire marketing and sales strategy is dirty data.
- Data Reshaping: Data Reshaping in R is the process of altering the data's row and column arrangements. When processing data in R, a data frame is used as the input.
- Statistical inference: The technique of inferring population characteristics from a sample of data is known as statistical inference.
- Sampling error: Sampling error is the term used by statisticians to describe the discrepancies between a sample and the population.
- Data manipulation: Data manipulation is the process of arranging data to make it more readable, visually appealing, or structured.
- Reproducibility: It is the degree to which a tool can deliver the same outcome when used again in similar situations. The terms repeatability and dependability are used interchangeably with the term reproducibility.
- Data bias: Data bias in finance refers to systematic errors or prejudices in financial datasets, which can lead to unfair or inaccurate outcomes. It can arise from various sources, such as historical, incomplete data collection, or algorithmic biases.
- Linear: The relationship between the independent and dependent variables is linear.

Check your Understanding

1. _____ is the process of preparing raw data for analysts to utilise in quick decision-making by cleaning, organising and changing it into the necessary format.
   a) Data Manipulation
   b) Data Wrangling
   c) Data Visualisation
   d) Data Bias
2. Data wrangling and ____ are terms that are frequently used interchangeably. Nevertheless, these are two completely distinct processes.
   a) Data discovery
   b) Data structuring
   c) Data Cleaning
   d) Data enriching

Notes

3. What is/are the use/uses of web scraping?
   a) Machine learning
   b) SEO monitoring
   c) Competition analysis
   d) All the above
4. Which web scraping entails reading the code and pulling out the pertinent information based on what the user needs?
   a) HTML parsing
   b) Automated extraction technique
   c) DOM Parsing
   d) Web scraping software
5. What is the main cause of the rise in copyright violations and terms of service violations?
   a) Data Wrangling
   b) Data Manipulation
   c) Web Scraping
   d) None of the above
6. _____ can result in inaccurate data recovery, ineffective personalisation, wasteful workflows, overcrowded storage systems and repetitive customer interactions.
   a) Duplicate Data
   b) Inaccurate Data
   c) Outdated Data
   d) Insecure Data
7. Which type of dirty data is worst data pollution?
   a) Incomplete data
   b) Inaccurate data
   c) Incorrect data
   d) Duplicate data
8. Which function is used to connect multiple vectors in R to produce data frames?
   a) cbind()
   b) rbind()
   c) merge()
   d) cast()
9. Which function is used to modify the shape of the data in several steps to get the desired shape?
   a) merge()
   b) melt()
   c) cbind()
   d) None of the above

Notes

10. Which inferential method is a projection of the population's outcome-producing process?
    a) Hypothesis testing
    b) Regression modelling
    c) Confidence interval
    d) Margin of error
11. Which type of probability is frequently used to compare a subset of data to the overall quantity of data obtained?
    a) Classical
    b) Subjective probability
    c) Relative frequency
    d) None of the above
12. What is the format of date in R?
    a) YYYY-MM-DD
    b) MM-DD-YYYY
    c) DD-MM-YYYY
    d) None of the above
13. The process of separating pertinent information from enormous quantities of textual material is referred to as _____.
    a) Information Retrieval
    b) Information Extraction
    c) Clustering
    d) Summarization
14. Which tool is used for summarization in text mining?
    a) Tropic tracking tool
    b) Rapid miner
    c) Intelligent miner
    d) Clear forest text
15. Which tool is used for cluster in text mining?
    a) Text Analyst
    b) Text finder
    c) Sentence ext tool
    d) Carrot
16. ____ arises when some research are more difficult to find than others.
    a) Citation bias
    b) Language bias
    c) Publication bias
    d) Location bias

Notes

17. ____ occurs when randomization while data collection is improperly achieved
    a) Sampling bias
    b) Automation bias
    c) Convergence bias
    d) Participation bias
18. In which machine learning algorithm mapping input and output data is the main objective?
    a) Supervised
    b) Unsupervised
    c) Reinforcement
    d) None of the above
19. What is/are the example/s of unsupervised machine learning algorithm?
    a) K-means clustering
    b) Apriori
    c) Eclat
    d) All the above
20. The ____ classifier uses independent Boolean variables as predictor variables. such as determining whether a word is used or not in a document.
    a) Gaussian
    b) Multinominal
    c) Bernoulli
    d) All the above

Exercise

1. Explain briefly what is Data wrangling?
2. What is the importance of Data Wrangling?
3. Analyse the challenges of Data wrangling.
4. Describe web scraping.
5. Explain concept of dirty data.
6. Analyse the concept of data manipulation.

Learning Activities

1. Describe machine learning algorithm with example.
2. What is bias in Data science? Explain its type.

Check your Understanding-Answers

1.	b	2.	c
3.	d	4.	a
5.	c	6.	a
7.	b	8.	a

Notes

9. b
10. b
11. c
12. a
13. b
14. a
15. d
16. d
17. a
18. a
19. d
20. c

Further Readings and Bibliography

1. Field Cady. The Data Science. 2017
2. William Vance. Data science: 3 Book in 1 – Beginner's Guide to learn the Realm of Data Science. 2020
3. Peter Bruce Andrew Bruce. Practical Statistics for Data Scientist. 2020
4. Reema Thareja. Data Science and Machine Learning using Python. 2022
5. Uma Maheshwari R Sujatha. Introduction to Data Science: Practical Approach with R and Python. 2021

Module - IV: Introduction to Data Science Tools

Learning Objectives

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

- Describe the tools and packages of data science
- Know the API
- Understand the types of visual comparison
- Describe the visual patterns
- Discuss the visual relationship
- Know type of visual proportion

Introduction

Data scientists utilise a variety of tools at various stages of the data science life cycle to process zettabytes and yottabytes of structured and/or unstructured data every day and to derive insightful knowledge from it.

The key benefit of these tools is that they do away with the requirement for complex programming languages to implement data science techniques. This is since these tools include some predefined algorithms, functions and graphical user interfaces (GUIs) that are user-friendly.

4.1 Data Science Tools and Packages

With the aid of various data processing techniques like statistics, computer science, predictive modelling and analysis and deep learning, data scientists can delve into complex, unstructured or structured data and process, extract and analyse it to uncover insightful information.

Software libraries or frameworks known as “data science packages” offer functions and tools to make data analysis, manipulation, visualisation and modelling easier. These tools are made to make it simpler for data scientists to use different analytical methods and work with data.

4.1.1 Languages of Data science

For data science, there are numerous programming languages available. Data scientists should study at least one language as it is a necessary tool for a variety of data science functions.

andProgramming Languages for Data Science

1. Python

   The most popular programming language for data science nowadays is Python. Python is currently the most widely used computer language for data science. It has been in use since 1991 and is a simple, open-source language. This versatile language is object-oriented. It also supports many paradigms, such as procedural, structured and functional programming.

2. JavaScript

   Data scientists also employ JavaScript, another object-oriented programming

Notes

language. There are currently hundreds of Java libraries available, each one addressing a different type of programming issue. Some unusual languages exist for developing dashboards and data visualisation.

This flexible language can handle numerous jobs at once. Everything from electronics to desktop and web programmes can be embedded with its help. Java is used by common processing systems like Hadoop. Additionally, it is one of those data science languages that is quick and simple to scale up for huge applications.

3. Scala

This sleek, contemporary programming language was developed in 2003, which is a lot more current. In the beginning, Scala was created to solve problems with Java. Web programming and machine learning are just two examples of its applications.

For managing massive data, it is also a scalable and efficient language. Scala allows concurrent and synchronised processing in modern organisations, as well as object-oriented and functional programming.

4. R

R is a powerful programming language made by statisticians. Statistical computing and graphics are two common uses of the open-source language and software. But it has a few data science applications and R has several helpful data science libraries.

5. SQL

With time, the computer language known as SQL—structural query language—has become more well-liked for handling data. Although SQL tables and queries are used for a variety of purposes, data scientists may find it helpful to have a basic understanding of them when interacting with database management systems. This domain-specific language makes storing, altering and retrieving data in relational databases exceedingly straightforward.

6. Julia

A computer language called Julia was developed primarily for high-performance computational science and rapid numerical analysis. It swiftly applies mathematical concepts like linear algebra. Additionally, it is a great language for working with matrices. Both front-end and back-end programming can be done using Julia and programmes can incorporate its API.

4.1.2 Data Science Tools

Almost every industry nowadays relies heavily on data science, whether it is for planning, future forecasting, or making business decisions. Everything fits under the technology and trends we're embracing right now.

In 2022, a world dominated by the digital era, where there is an abundance of data and variety of tools is used and approaches to become resourceful for various goals. The only popular technology you would discuss is "Data Science".

Notes

Data Science Tools

An individual must be proficient in one or more programming languages and a variety of tools to execute specified tasks. Even if you dig deep, there are over 5,24,000 jobs accessible globally and more than 38,000 positions available in India as of right now.

These facts indicate that there is an expanding need for data science specialists in practically every industry, thus it is imperative to keep up with the most recent techniques and technology.

1. SAS

   There are various categories within data science and "data visualisation" is one of them. SAS is the name of a tool that ought to be first in the "statistical" category while you're working with visualisation. It is used to produce and exhibit symmetric analytics charts and aids in data management.

   It uses statistical modelling and the SAS programming language to achieve this. Learn this tool if you're interested in this subject or trying to break into data science because most businesses utilise comparable metrics and will expect you to be familiar with these kinds of tools. To master these tools and technologies, you need to have a fundamental foundation of data science.

   Other than the fact that it is among the priciest software in the market, only large-cap companies would require you to be familiar with this tool.

2. Microsoft Power BI

   The most effective tool you need to be familiar with when working on data visualisation. It provides to provide insight into any provided data as a cloud-based analysis service, which aids in business decision-making. It has the capacity to offer a thorough analytical setting for keeping track of reports from many angles.

   Its "Ease of Usage" is the main factor making it popular with data scientists and this makes it easier for people to use it for data visualisation.

3. BigML

   Techniques like data clustering, classification, anomaly detection, time-series forecasting, etc. that use ML algorithms might be another specialist tool used in data science for predictive modelling.

   In addition, it provides a user-friendly, cloud-based GUI environment that can be utilised for product innovation, risk analysis and sales forecasting. Today, BigML is being used by more than 50,000 people and it has a large international user base.

Notes

The finest part of BigML is that it enables customers to build their own private dashboards and when all information is retrieved through its API, they offer improved security by enabling HTTPS for efficient data and communication flow.

4. Knime

Konstanz Information Miner, sometimes known as KIM, is an open-source data analysis tool created in Java and based on Eclipse. Additionally, it presents the idea of a modular data pipeline that enables data mining. Perhaps this instrument is intended for data analysis, thus one should understand the fundamentals of that process.

KNIME also provides other services like data modelling, data preparation, visualisation, etc. It also provides data sets that have been combined, converted and filtered in data science, which makes it one of the top data science tools you should be familiar with.

5. Tableau

How can avoid Tableau when discussing popular data science tools? one of Tableau's most well-liked and perhaps most extensively used data visualisation products. This technology is well designed to support both business intelligence and data science.

Because of its simplicity, it enables non-technical persons to design their own customised dashboards and contributes to the creation of basic yet elegant data that is simple enough for professionals of any level (technical and non-technical). These elements make using this tool in projects more enticing for data science experts.

6. TensorFlow

Because "TensorFlow" is a popular tool among experts and data scientists, you may have heard of it. Being an open-source platform, it allows users to design data flow graphs where the graph's nodes prominently depict mathematical and statistical operations and the graph itself represents multidimensional arrays (data) that exchange information between them.

Due to this approach, ML can be seen as a connected graph of operations. Additionally, they are designed to run on a variety of platforms while aligning the GPU, CPU and TPU without the need for repetitive coding, which ultimately implies that anyone may use this tool to increase efficiency. In addition, it enables users to keep track of the training process and all the evaluation data.

7. Snowflake

Today, "Data Warehousing" is one of the most crucial subfields of "Data Science," and the snowflake, which is built on SQL for the cloud, is the ideal instrument to carry out this operation. The nicest thing about this is that it provides unrestricted flexibility and effectiveness, which is difficult to achieve without a big data platform.

In addition, it provides certain other advantages that increase its dependability for use as a data warehousing tool. A few of these advantages are:

- No one else can access the data that is kept there; rather, the user must access it using a Snowflake SQL query to view the data.
- It invests all its resources in building a virtual warehouse so that it can process and protect data even more quickly. This makes it possible to replicate the data across the cloud in the event of a breakdown (to maintain a successful business operation).
- A user can only access a specific amount of data for free, therefore each time they want to view the data (or any piece), they must pay a certain fee.

Notes

4.1.3 Data Science Packages

A package in Python is a grouping of modules. Usually, modules that connect to one another are packed together. A programme can import an external package and utilise its modules if it requires a module from it.

Many important tasks are streamlined by Python packages, including data analysis and visualisation, machine learning model construction, web data collection of unstructured data and effective processing of picture and text data. The top 10 Python packages for data scientists are listed below.

1. TensorFlow: One of the most popular machine learning libraries is TensorFlow and for good reason. It specialises in exploiting data flow graphs for numerical calculation. It functions similarly to a computational library when creating new algorithms with several tensor operations.
2. NumPy: The main Python tool for scientific computing is called NumPy. It combines the simplicity and adaptability of Python with the speed of C and Fortran-family programming languages. For a range of general-purpose programming tasks, it is a useful P
3. SciPy: Modules for optimisation, linear algebra, integration and statistics are included in the SciPy library. It is a sizable collection of data science tools, primarily in the fields of science, technology and mathematics. The fundamental data structure of SciPy is a NumPy array and it includes modules for many frequently performed operations in scientific programming.
4. Pandas: Pandas is a Python machine learning package that offers high-level data structures and a variety of analytical tools. It is regarded as a quick, effective and simple-to-use tool for data manipulation and analysis. It functions with data frame objects, which are special two-dimensional data structures.
5. Matplotlib: A Python 2D plotting toolkit called Matplotlib makes it simple to create cross-platform graphs and charts. Basic graphs including line plots, histograms, scatter plots, bar charts and pie charts are created using it.
6. Keras: Keras is designed for quick experiments. It can function while atop other frameworks. It offers a simpler method for expressing neural networks. Common neural network building pieces like layers, objectives, activation functions and optimizers are widely implemented in Keras.
7. SciKit-Learn: Because Scikit-Learn has such a low barrier to entry, even business-side personnel can use it. It is regarded as one of the top libraries for handling complicated data. Numerous algorithms are included for carrying out common machine learning and data mining tasks as dimensionality reduction, classification, regression, clustering and model selection.
8. PyTorch: The biggest machine learning library, PyTorch, enables programmers to generate dynamic computation graphs, execute tensor computations with GPU acceleration and compute gradients automatically. On a tape-based automatic grading system, it constructs dynamic neural networks.
9. Caffe: Convolutional Architecture for Fast Feature Embedding, or Caffe. It is one of the quickest convolutional network implementations, which makes it perfect for image recognition. The processing of images by Caffe is quite amazing.
10. Theano: You may define, optimise and effectively evaluate mathematical formulas involving multi-dimensional arrays using the Theano Python package. It is among the

Notes

earliest deep learning software libraries available as open source. The greatest option for quick calculations.

4.1.4 APIs

You will inevitably run into references to something called a "API" if you work in the tech industry. Simply said, you can't skip it since you will inevitably hear it again if you do. There are APIs practically everywhere.

The web is among the most widely used platforms for APIs. If you've spent any time online, you've used APIs. These services all utilise API in the background, whether they are used to process payments online, share content on social media, or show a list of tweets through a social handle.

APIs are frequently used by developers to add various functionalities to their creations. To integrate advanced features, companies can make a simple API call within their product rather than having to write the code themselves.

Basic Elements of an API:

An API has three primary features:

1. Access: is the user or who has the authority to request information or services?
2. Request: Exactly what information or service is being requested? A Request has two main parts:
   - ❖ Methods: i.e., the questions you can ask, assuming you have access (it also defines the type of responses available).
   - ❖ Parameters: you can include more information in the query or response.
3. Response: due to your request, the information or service.

Categories of API

1. Web-based system

   An interface to a web server or a web browser is referred to as a web API. The creation of web apps makes extensive use of these APIs. Both the server end and the client end of these APIs are functional. Web-based APIs are offered by businesses like Google, Amazon and eBay.

   The Twitter REST API, Facebook Graph API, Amazon S3 REST API and other popular web-based APIs are some examples.

2. Operating system

   Applications for Windows or Mac can be made using a variety of OS-based APIs that provide functionality for different OS capabilities. Some of the examples of OS based API are Cocoa, Carbon, WinAPI, etc.

3. Database system

   The API calls to the database are used to interact with most of the database. These APIs are designed to send the desired data in a format that the client making the request may comprehend. Some popular examples are Drupal 7 Database API, Drupal 8 Database API, Django API.

4. Hardware System

   Access to the various hardware parts of a system is made possible by these APIs.

Notes

They are necessary for connecting to the hardware and establishing communication. Because of this, it enables a variety of activities, including the gathering of sensor data and even the presentation of information on your screens. Some other examples of Hardware APIs are: QUANT Electronic, WareNet CheckWare,OpenVX Hardware Acceleration, CubeSensore, etc.

5 APIs every Data Scientists should know

1. Facebook API

   An interface to the vast amount of data generated daily is provided by the Facebook API. Massive amounts of data are produced by the countless posts, comments and shares in different groups and sites. And there are numerous potentials for crowd analysis given the vast amount of public data. Additionally, it is exceedingly straightforward to extract data using R and Python with the Facebook Graph API.

2. Google Map API

   One of the often-utilised APIs is the Google Map API. Its uses range from inclusion in a cab service software to the well-known Pokemon Go. You can get access to all the data, including route information, route distances and position coordinates. The interesting aspect is that you can also add a distance feature to your datasets using this API.

3. Twitter API

   Twitter data can be accessible using the Twitter API, just like Facebook Graph API. You have access to all the information, including tweets from any user, tweets that contain a specific term or even a group of related terms, tweets posted on the issue during a specific time frame, etc. Twitter data is a fantastic resource for activities like sentiment analysis and opinion mining.

4. IBM Watson API

   With just a few lines of code, IBM Watson's APIs can handle a variety of difficult tasks like text-to-speech, speech-to-text, personality insights, visual recognition and tone analysis. This group of APIs is distinct from the others mentioned thus far since they offer services for modifying and drawing conclusions from the data.

5. Quandl API

   With Quandl, you may access the time series data for numerous stocks for the chosen time frame. The Quandl API is simple to set up and offers a tremendous resource for tasks like stock price prediction and stock profiling, among others.

4.2 Types of Visual Comparisons

In this instance, data visualisation is useful. Many businesses prefer using dashboards for data visualisation to share information and evaluate it in order to make it more comprehensible and accessible.

4.2.1 Types of Visuals

Data visualisation is the process of presenting information and data graphically. By including visual components like charts, graphs and maps, data visualisation tools provide a simple method for identifying trends, outliers and patterns in data. Furthermore, it provides employees or business owners with a terrific approach to convey information to non-technical people without creating misunderstanding.

Notes

Visualising data is enjoyable. It transforms important data discoveries and analytics research into fun pictures that you can pinch, twitch and manipulate. To stay competitive, it aids businesses in identifying trends, developing long-term business plans and making crucial decisions more quickly.

Visual comparison methods are essential for data analysis and interpretation in data science. They aid in extracting insights from the data and spotting patterns, trends and linkages. Here are a few examples of typical visual comparisons in data science:

1. Bar Charts: Comparing discrete values or categorical data with bar charts is helpful. Rectangular bars of varying widths or heights are used to display the data, with each bar's length corresponding to the value it stands for. When comparing data between different categories or groups, bar charts are useful.
2. Line Charts: The visualisation of trends and patterns across time can be done with line charts. They are produced by drawing straight lines linking data points on a Cartesian coordinate system. To compare changes or variations in a variable or several variables over a continuous period, line charts are frequently employed.
3. Scatter Plots: Two continuous variables are compared using scatter plots. One variable is represented on the x-axis and the other on the y-axis and each data point is depicted as a dot on the graph. Scatter plots enable the detection of correlations, clusters, outliers and interactions between the variables.
4. Heatmaps: When comparing data in a matrix or grid format, heatmaps are useful. To show the size of values within the grid, they employ colour coding. When analysing enormous datasets, heatmaps make it possible to visually identify patterns, clusters, or changes in the data.
5. Box Plots: When examining the distributions of continuous variables among various categories or groups, box plots, sometimes referred to as box-and-whisker plots, are helpful. They show the data's minimum, maximum, median and quartiles, giving a clear overview of the distribution and making it easier to compare groups.
6. Trees: Trees use nested rectangles to show hierarchical data structures. They make it possible to compare the percentages or contributions of various categories within a hierarchy. A visual representation of the hierarchical structure and relative sizes of the categories is provided by the size and colour of each rectangle, which can represent different attributes or values.

These are only a few examples of visual comparison methods in data science. The type of data, the variables being compared and the precise objectives of the study all influence the technique choice..

4.2.2 Tables

The table is a crucial sort of object for displaying data collections. A table can be seen in one of two ways: as a collection of rows that each include all the information about a single entry in a data set, or as a collection of named columns that each describe a specific aspect of all entries in a data collection.

To use tables, import all the module called data science, a module created for this text.

```
from datascience import *
```

Empty tables can be created using the Table function. An empty table is useful because it can be extended to contain new rows and columns.

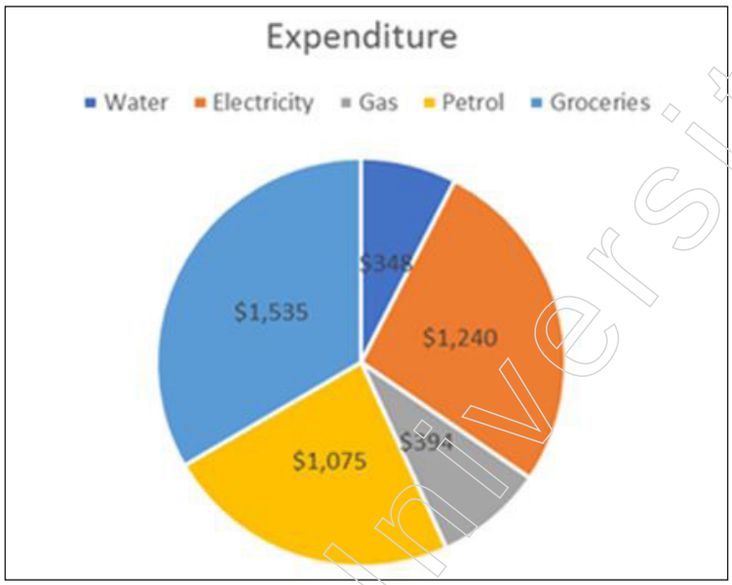
Notes

Table()

The with_columns method on a table constructs a new table with additional labelled columns. Each column of a table is an array. To add one new column to a table, call with_columns with a label and an array. (The with_column method can be used with the same effect.)

4.2.3 Pie Charts

One of the often-used components in Excel worksheets to display official data is the pie chart. A pie chart is described as a circular chart with numerous sections, each of which represents the percentage that each value contributed to the overall value. Each slice of the pie chart adds a certain percentage to the overall value, with the pie (or circle) representing the whole value, or 100%.



Pie Chart

Typically, data that can be added together or that only has one data series (all the data points are positive) is represented with pie charts. As an example, consider how to represent the expense list of a xyz company in the example above using a pie chart. The pie chart (also called as slices of a pie) displays 5 divisions because the company has 5 separate features.

Advantages of Pie Charts

1. Easy to create: Although it's simple to generate most Excel charts, it is assured that Pie charts are the simplest. Pie charts allow you to easily customise and style your chart because, for the most part, the default choices work well.
2. Easy to read: If you merely have a few data points, pie charts are simple to read and to see.
3. Management is obsessed with Pie Charts: According to a survey, managers and clients adore pie charts in formal presentations.

Disadvantages of Pie Charts

1. Pie charts are helpful when there are fewer data points, but they can get complicated when there are more data points.

Notes

2. Pie charts are designed to offer you with a snapshot of the numbers at a specific point in time; they cannot be utilised to display a trend.
3. Pie charts cannot display many kinds of data values.
4. Pie Diagrams Although the differences between the data points are small, you could find it difficult to interpret the pie chart visually.
5. Pie charts require more room and provide less information.

Types of Pie Charts

To depict your PIE charts in Excel, there are various forms available. Go to the charts group, select the Pie chart option and a drop-down menu will appear so you may access its numerous forms. It will display every type of Pie chart that Microsoft Excel offers.

1. 2D Chart

2D or 2-Dimensional charts are further divided into 3 types which are as follows:

Pie

The other 2D pie is Normal Pie. It is used to show the contribution of each point to total value.

Exploded Pie

Exploded Pie highlights various data values while displaying how each value contributes to the overall value. Additionally, you may "explode" a standard pie by clicking on it, choosing a slice and then dragging it away from the centre.

PIE of PIE Chart

When adding values to the second pie from the primary pie, a pie of pie chart is employed. Using this chart, you can emphasise more values or make minor percentages easier to understand.

Bar of PIE Chart

The main pie's values are taken from the bar of pie chart and combined into a stacked bar. Using this chart, you can emphasise more values or make minor percentages easier to understand.

The only difference between the Bar of Pie chart and the Pie of the Pie chart, as you can see in the example, is that a sub bar will be formed in place of a sub pie.

2. 3D PIE Chart

The differences between 2D and 3D pie charts are minimal. Apart from appearance, both are nearly identical. This is because, in contrast to 2D charts, which only have length and breadth, 3D charts also contain depth.

4.2.4 Bar Charts

In a bar chart, also referred to as a bar graph, categorical data is displayed using rectangular bars with heights that match to the values they represent. The value scale is shown on one axis of the chart in this instance and the categories are plotted on the other. Data comparisons can be done right away because the bars are all the same width.

Notes

Bar Chart

Types of Bar Charts

1. Pareto charts

   The frequency charts are ranked from highest to lowest.

2. Column or Vertical Bar Charts

   The values of each category are represented by the height of the bar, which is displayed along the horizontal axis with the categories. These graphs work really well for showing data sets over time.

3. Stacked Bar Chart

   Bars representing various groupings are stacked on top of one another. The height of the resulting bar displays the summation of the groups' results.

4. Grouped Bar Chart

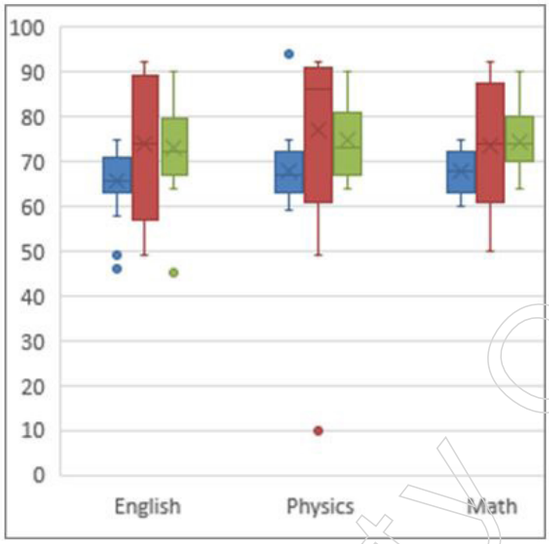
   There are two or more coloured bars for each classification group to signify a specific grouping.

4.2.5 Box and Whisker Plots

A box and whisker plot, which also emphasises the mean and outliers, shows the distribution of the data into quartiles. There may be vertical "whiskers"—lines—on the boxes. Any point outside of the whiskers or lines delineating the variability between the upper and lower quartiles is regarded as an outlier.

In statistical analysis, box and whisker plots are most frequently utilised. For instance, you may compare test results from teachers or medical trials using a box and whisker plot.

Notes



Box chart

4.3 Types of Visual Patterns

The three-dimensional representation of a bag of words has been discriminatively abstracted using visual patterns, or high-order combinations of visual words. The current visual pattern mining approaches, however, tie words from distinct objects/depths wrongly into the same pattern to degrade the mining precision since they are based on poorly posed 2D photographic concurrences of visual words rather than their real-world 3D concurrences.

However, it is still unclear how to create a condensed yet discriminative image representation from the patterns extracted, even though this is much desired for many new applications like mobile visual search.

4.3.1 Line Charts

In a line chart, which is sometimes called as a line plot or line graph, points are connected using line segments that run from left to right to depict changes in value. Along the continuous development of the horizontal axis, which is often that of time, the vertical axis displays data for an interest metric.

Line Chart

Notes

The exchange rate between two fictitious currencies is depicted in the line graph above for a six-month period. A series of points connects the daily exchange rates as time moves from left to right. The rate increased from approximately 0.75 to 0.78 between March and early April, then progressively decreased to approximately 0.765 in late May and early June, according to the line's general slope and vertical positions.

When you should use a line chart

You require a variable to represent continuous values with a regular measurement interval on the horizontal axis. This variable frequently produces an observation once every minute, hour, day, week, or month. Instead of being an intrinsic property of the data, the analyst will typically need to decide on the interval size, or bin, for the data.

For each point that lies within an interval specified by the horizontal axis variable, you must provide the value of a second numerical variable on the vertical axis. This will frequently be a statistical summary, such as the sum or average value over all events within each bin.

A single line chart can also have many lines plotted to compare the trends between different series. This is frequently used to track how the data is distributed among various subgroups. The line chart has a unique use case made possible by the ability to plot many lines in situations when it might not often be chosen.

4.3.2 Area Charts

An area chart is a type of graph used to show changes in amounts over time that combines a line chart with a bar chart. In that line segments are used to link data points, it resembles a line graph. The region below the line, however, is filled in or shaded. A chart with layers is created by plotting additional numbers below the lines and shading them in a different colour.

The earliest area charts were created by William Playfair and were included in his book from 1786. Several time-series graphs in this book, The Commercial and Political Atlas, featured statistics on imports and exports as well as national debt. Playfair also created the pie, bar and line charts.

Types of Area Charts

1. Overlapping Area Chart

   This graph compares values for the various categories and demonstrates how the data overlap. For instance, a graph of iPhone sales can indicate a peak in one type of phone's sales followed by a decline as another model is introduced. Although there will be a crossover area, the chart will clearly show sales peaks and then declines. The transparent shading draws attention to the transition.

2. Stacked Area Chart

   Most people refer to this chart as an area chart, so we'll use that term throughout this essay. A stacked area chart uses solid colours or patterns, but an overlapping area chart shows crossing using translucent shading. When tracking overall value and the distribution of that amount over groups, a stacked chart is utilised. To compare the performance of one group to another, it is simple to compare the height of the stacked area.

Notes

When to Use an Area Chart

The best way to display different patterns over time is via an area chart. It is best used when:

- There is data expressed as a total
- There are time periods to compare
- The point of the chart is to communicate an overall trend, not individual values
- There are multiple data series with part-to-whole relationships, or a cumulative series of values.

However, the graph is deceptive because it gives the impression that Boston has the most sales. This figure, meanwhile, only represents a portion of the total volume; in terms of sales among consumers aged 60 to 69, Seattle leads the nation. The major issue with area charts is that they can be deceptive and challenging to understand.

When Not to Use an Area Chart

Area charts are only useful for a specific set of circumstances. There are far better alternatives for most other data sets:

- An area chart is not the best choice if you want to illustrate how values vary across several different categories. Consider utilising a bar chart, column chart, or split bar chart as an alternative.
- If the data total is not significant, an area chart may be deceptive and exclude vital information. A line chart, like the electronics sales chart in the example above, will probably be simpler to read and comprehend.
- A line chart is preferred when there are slight variations in the values. This is since its Y-axis can be expanded to display even minute variations and need not begin at zero.
- A line chart or column chart is preferable and labelling is ideal, if the chart just shows one number across time or there are only a few dates to plot. A simpler way to read a stacked column chart is when there are ten or fewer dates.
- A line chart presents the data more effectively when comparing the magnitude of various shares.

4.3.3 Scatter Charts

A scatter chart, also known as a scatter plot, is a type of graph that shows the relationship between two variables. They are very effective chart types that let readers to see relationships or trends right away that are difficult to see in practically any other format. Modern scatter charts are based on René Descartes' 17th-century cartesian coordinates system, albeit their exact origins are unknown. Scatter plots are often used in science, with most of them appearing in scientific journals and publications.

How Does a Scatter Chart Work?

A scatterplot has an X and a Y axis, just like the majority of other graph and chart types. The dependent variable is represented by the vertical Y while the independent variable is represented by the horizontal X. A mark or dot is then placed at the location that symbolises the junction of the two coordinates after creating an even scale on both axes.

There are other patterns to be found within a scatter chart:

Notes

- Linear or nonlinear: Through the data points, a linear—straight—correlation can be created, but a non-linear correlation may reveal a curved link.
- Weak or strong: The dots will be more closely spaced apart the stronger the link. A poor association will result in more dispersed data points.

Many scatter plots make use of trend lines to vividly illustrate these connections and trends. On the graph, a trend line is drawn to highlight the strength and direction of the trend.

When to Use Scatter Charts

Aside from scientific studies, there are a few times when businesses may decide to use a scatter chart:

- ❖ To identify anomalies
- ❖ To see how one variable affects another
- ❖ To see a correlation, pattern, trend, or relationship

Benefits of Scatter Charts

Scatter charts have multiple benefits and advantages.

1. Clearly Shows Relationships

   This graph, in my opinion, illustrates correlations between two variables the best. It demonstrates a pattern or trend over the entire data collection, not just a link between two data points.

2. Easy to Create and Understand

   Scatter plots are simple to understand, maybe because of their widespread use. Their goal is clear and their information is simple to understand. Furthermore, scatter charts are easy to build for people who want to use them.

3. The Range of Data Can be Determined

   Scatter plots make it possible to observe the greatest and minimum values, which is crucial for comprehending the complete collection of data. Outliers, though, might be confusing.

Disadvantages of Scatter Charts

1. Can Have Too Much Data

   Patterns are difficult to spot in a scatter plot chart that has been over-plotted since it resembles a huge blob. Therefore, even if a chart requires sufficient data to show a noticeable correlation or pattern, there comes a point where additional data is no longer as useful.

2. No Relationships

   Data may occasionally seem to show a pattern or correlation. But despite the appearance that height and cat ownership are related, they most likely are not.

3. Correlation Does Not Equal Causation

   Continually keep in mind that correlation does not imply causation. There is no guarantee that one thing causes another just because there is a correlation. Although it may appear that tall people own more cats, it is doubtful that being tall contributes to cat ownership.

Notes

4.3.4 Cluster Charts

The unlabeled dataset is grouped using the machine learning approach of clustering or cluster analysis. One definition of it is "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

To do this, it separates the unlabelled dataset into groups based on whether certain associated patterns, such as shape, size, colour, behaviour, etc., are present. It employs unsupervised learning, so the algorithm receives no supervision and operates on an unlabelled dataset.

The clustering technique can be widely used in various tasks. Some most common uses of this technique are:

- Market Segmentation
- Statistical data analysis
- Social network analysis
- Image segmentation
- Anomaly detection, etc.

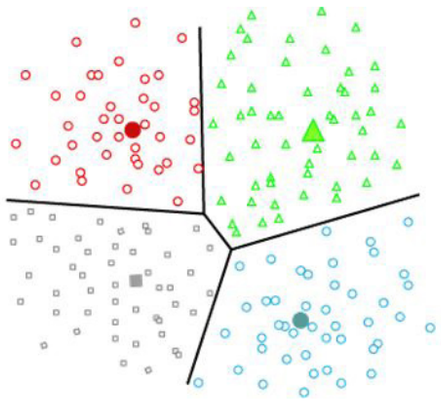
Types of Clustering Methods

The two main types of clustering are Hard Clustering (where a data point can only belong to one group) and Soft Clustering (where a data point can also belong to another group). However, there are several different Clustering techniques available. The primary clustering techniques in machine learning are listed below:

1. Partitioning Clustering

   It is a type of clustering in which the data are grouped non-hierarchically. The centroid-based technique is another name for it. The most often used instance of partitioning clustering is the K-Means Clustering technique.

   In this type, the dataset is split into a collection of k pre-defined groups, where K refers to the number of groups. When compared to another cluster centroid, the distance between the data points of one cluster are separated by the cluster centre in the smallest possible amount of space.



Partitioning Cluster

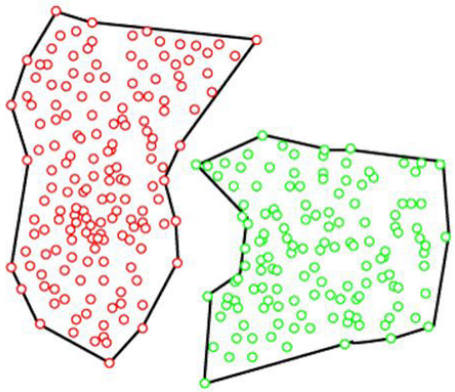
2. Density-Based Clustering

   The highly dense regions are joined into clusters using the density-based clustering method and if the dense region can be connected, the distributions can take any

Notes

shape. This programme accomplishes that by finding several clusters in the dataset and joining the regions with dense population into clusters. Sparser regions separate the dense areas in data space from one another.

If the dataset comprises a wide range of densities and high dimensions, these algorithms may struggle to cluster the data points.

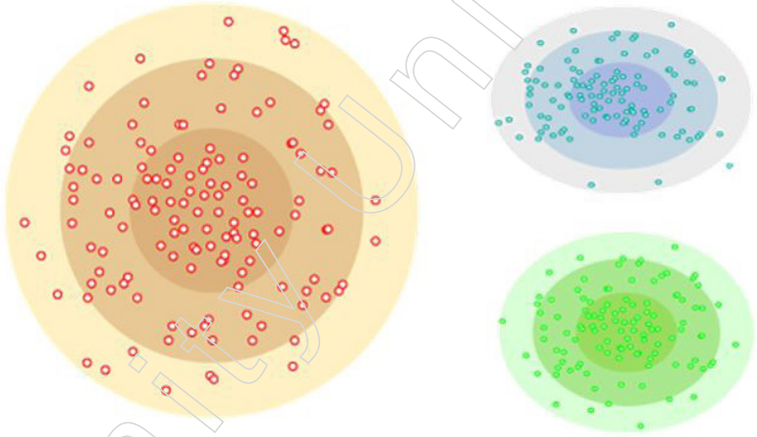


Density based clustering

3. Distribution Model-Based Clustering

The chance that a dataset conforms to a certain distribution is used to partition the data in the distribution model-based clustering approach. The categorization is carried out by usually assuming some distributions. Statistical Distribution.

The example of this type is the Expectation-Maximization Clustering algorithm that uses Gaussian Mixture Models (GMM).



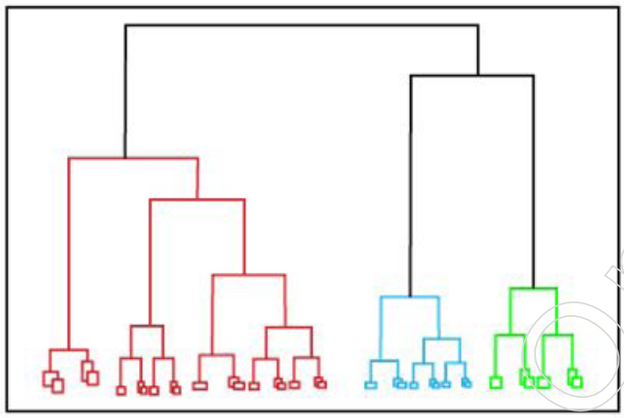
Distribution model-based clustering

4. Hierarchical Clustering

Hierarchical clustering can be used instead of partitioned clustering because it is not necessary to pre-specify the number of clusters to be created. With this technique, the dataset is divided into clusters to create a dendrogram, which resembles a tree.

By removing the appropriate amount of the tree, it is possible to choose the observations or any number of clusters. The Agglomerative Hierarchical algorithm is the most prevalent example of this approach.

Notes



Hierarchical clustering

5. Fuzzy Clustering

   A data object may be a member of more than one group or cluster when using a soft method called fuzzy clustering. Each dataset contains a set of membership coefficients that vary depending on how much of a cluster a dataset a part of is. This sort of clustering is exemplified by the fuzzy c-means method, which is also referred to as fuzzy k-means algorithm.

Applications of Clustering

Below are some commonly known applications of clustering technique in Machine Learning:

1. In Identification of Cancer Cells: To find cancerous cells, clustering approaches are widely used. From the carcinogenic and non-cancerous data sets, it separates them into various groups.
2. In Search Engines: Search engines also employ the clustering technique. The search outcome is presented based on the object that is closest to the search query. This is achieved by grouping together similar data objects in a different group from the other, divergent objects. How accurate the results of a query are depending on how well the clustering algorithm performs.
3. Customer Segmentation: Customers are divided into groups depending on their choices and preferences in market research.
4. In Biology: The image recognition technology is employed in the biology stream to categorise various species of plants and animals.
5. In Land Use: To locate areas of similar land use in the GIS information, the clustering technique is applied. Finding the best use for a piece of land, or the purpose for which it is most appropriate, can be quite helpful.

4.3.5 Density Charts

Also known as a Kernel Density Plot or Density Trace Graph.

Because the number of bins utilised (each bar in a conventional histogram) has no bearing on density plots' ability to determine the distribution shape, density chart offer an advantage over histograms in this regard.

A histogram with only four bins would not yield a sufficiently identifiable distribution shape as would a histogram with twenty bins. However, this is not a problem when using

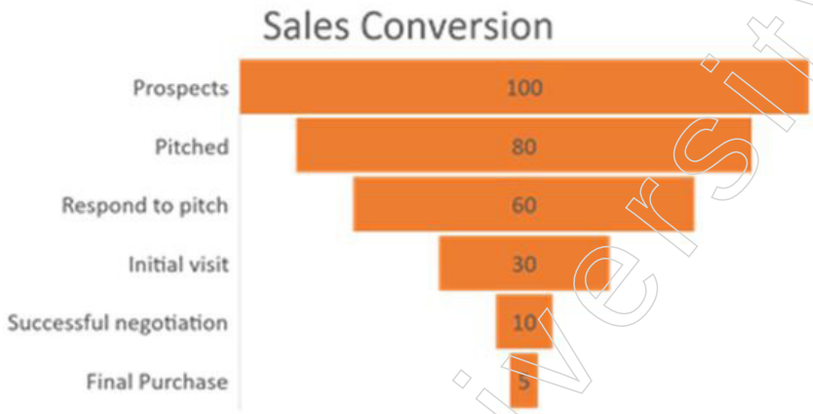
Notes

density plots. The histogram can be thought of as being expanded upon by a density chart. In contrast to the histogram, the density chart can eliminate noise and smooth the value distribution.

Due to the fact that the shape of a probability distribution can be determined using such density charts, which are unaffected by the number of bins, it is possible to find an appropriate distribution to use to model the statistical characteristics of a population given a sample from that population.

4.3.6 Funnel Charts

A funnel chart is a type of graphic that shows how data flows through a process. In a funnel diagram, the value of the dependent variable decreases as the process progresses. Many processes, including recruitment, order fulfilment and sales funnels are represented by funnel charts.



Funnel chart

100 possible clients are at the top of the funnel. Only 60 of these customers respond when the real estate company begins pushing their items to them. The 60 clients then go to the following stages of the sales funnel.

Only five consumers made a purchase from the company of real estate. The sales process is illustrated in the funnel diagram below. Each bar's length corresponds to the quantity of consumers at each stage of the sales process.

When to Use a Funnel Chart?

A funnel chart is a highly accurate representation that is only suitable for a specific kind of data. funnel charts suit the data well if it has the following characteristics:

- The data travels through a series of steps and there are at least three steps in all.
- Each level sees a decrease in the data. There are more objects in the first stage than the second. More items are in the second than the third and so on.
- The data are numerical and each level clearly shows a reduction in data.

What are the Main Uses of Funnel Charts?

The best way to show how a variable (in the example above, the number of consumers) changes throughout the course of a transaction is to use a funnel chart. This could either be a sales process or a selection process.

Notes

In general, funnel charts are used for various purposes:

1. Visualize Customer Drop Out

   There are several customers present at the start of a normal sales transaction. At various moments during the sales process, some consumers leave. A funnel diagram shows the percentage of customers who move on to the next step of the sales process. In other words, the chart also represents what percentage of the customers have dropped from the sales process.

2. Visualize a Selection Process

   Additionally, funnel diagrams can depict a selection procedure, such as an interview or a contest. Consider, for instance, that 100 people applied for a single position. Only 60 of them made it past the resumption inspection. Only forty of them made it past the phone interview.

   This interviewing procedure is shown in the funnel diagram below. It displays the number of applicants who attended each round of the hiring process.

3. Visualize Bottlenecks in a Process

   For locating a bottleneck in a business process, funnel charts are quite helpful. The funnel diagram below, for instance, depicts a potential email marketing flow. Only 50% of customers who have opened the email are clicking on the link within it in the step below. This may be due to an issue with the email.

4. Visualize Order-Fulfilment

   A funnel chart is ideal for use in an electronics store when evaluating order fulfilment. The number of initial orders is shown at the top of the graph. The following bars show how many orders were submitted, approved, dispatched and delivered.

What Are the Common Variations of a Funnel Chart?

Funnel charts can be represented using various formats:

1. Inverted Triangles

   Inverted triangles are a common visual representation of funnel charts. The problem with showing a funnel chart as an inverted triangle is that users are frequently persuaded to think that the area of each segment corresponds to the actual value of the data. The length of each segment instead of its area represents the facts.

2. Set of Diminishing Bars

   A series of bars with each bar having a shorter length than the one above it would be a more accurate portrayal of a funnel chart. It is evident from this depiction that each bar's length corresponds to a particular data variable's value.

What Are the Best Practices when Using a Funnel Chart?

1. Use Clear Annotations

   An effective technique for visualisation is a funnel chart. It might be improved by using clear annotations. Use labels that are clear and succinct. The chart will get cluttered and the viewer will become confused if the labels contain too much information.

   In a funnel chart, it is a good idea to include the percentage values in addition to the raw data. The percentage drop in each stage of the sales or business process is better understood as a result.

Notes

2. Use an Effective Colour Scheme

   To improve the clarity of a funnel chart, use a colour scheme for the bars. A frequent practise is to give each bar a different hue. Viewers could become perplexed if a gradient changes gradually or uses the same colour.

3. Number of Stages

   Funnel charts are appropriate for data that goes through at least three stages. A pie chart or bar chart is preferable when there are fewer stages involved.

What Are the Advantages of Funnel Charts?

1. Funnel Charts are Easy to Visualize
2. Gives Emphasis to Processes
3. Help to Identify Bottlenecks

What Are the Disadvantages of Funnel Charts?

1. Only Represents a Single Variable
2. Not Suited for Data Analysis

4.4 Types of Visual: Changes in Prices

4.4.1 Candlestick Charts

A type of financial graph known as a candlestick chart frequently shows movements in the price of commodities, stocks, or derivatives. It resembles a candlestick because of its vertical rectangle and wicks at the top and bottom. The open and closed prices are shown at the top and bottom of the candlestick. At the top and bottom of the wick, respectively, the price is shown.

Since a Japanese guy by the name of Homma created them in the 1700s, candlestick charts have been in use. He understood that there was a connection between the price of rice, supply and demand and the feelings of the rice dealers.

The open, the high, the close and the low were represented by each candlestick on the chart he created, which also included the four other dimensions in a trading period. After the West caught up with him a century later, the rest is history.

For traders and investors wishing to do thorough research before making an investment, these charts that show market volatility and trends can be a useful resource. Any trader's plan must include employing candlestick technical analysis to decide when to enter and quit trades.

How Does a Candlestick Chart Work?

Candlestick charts show a range of information:

- Open price
- Close price
- Highest price
- Lowest buy price
- Patterns and trends in share prices
- Emotions of trades

Notes

The rectangular candlestick, which displays the open and closing price of the shares and stocks (or rice), is referred to as the “real body”. The wick, also known as the “shadow,” which extends from the real body’s top and bottom displays the day’s high and low-price points.

The close price was lower than the open price if the true body is black or filled in. The close price is greater than the starting price if the genuine body is open.

The wicks also give significant information. Short upper wicks on down candles indicate that the open price was near the high price, while short upper wicks on up candles indicate that the closing price was near the high price.

Another pattern is known as a “doji,” in which the candlestick’s body has almost entirely vanished and just a cross is visible. Three distinct doji subtypes exist:

- Dragonfly: a long bottom wick and a short top wick where there is no body. This demonstrates that consumers prefer higher costs over reduced prices.
- Gravestone: when a higher price is rejected in favour of a lower price as shown by the long top wick and short bottom wick.
- Long-legged: has two lengthy wicks; it displays ambiguity and no discernible pattern.

The dragonfly and gravestone doji are both related to a hammer and an inverted hammer. In contrast to dojis, who have no body at all, they do, however, have miniature genuine bodies.

Numerous candlesticks on a single chart reflect a range of patterns. Bullish or bearish terms are used to describe these trends and patterns. Bearish patterns signal that a price is likely to decline and bullish patterns reflect tendencies that suggest a possibility the price will climb. These patterns demonstrate movement trends but are not perfect forecasts.

Trading is mostly governed by emotion, despite certain quantitative components. Candlestick charts, when viewed holistically, exhibit this feeling. The candlestick’s appearance is affected by how the open, close, high and low are related.

Candlestick Chart Patterns

The smaller the wick and the longer the body, the stronger the bullish movement or price gain. The smaller the body and the larger the wick, the more bearishly, or downwardly, the price is moving.

1. Bearish Engulfing Pattern

   In this pattern, sellers outweigh buyers, indicating an upward tendency. Typically, there is a lengthy, strong real body engulfing a little, open real body, indicating a seller’s market and the possibility that the price will continue to decline.

2. Bullish Engulfing Pattern

   When there are more buyers than vendors, this pattern is seen. A little, solid body is engulfed by a lengthy, open genuine body, indicating that the price may increase.

3. Bearish Evening Star

   This is a topping pattern in which the last candle opens below the small genuine body of the previous day. The last candle shuts firmly in the candle’s actual body from the previous two days. This pattern reveals that buyers procrastinate before sellers seize the initiative. It suggests that there may be a rise in sales.

Notes

4. Bearish Harami

   The actual body from the previous day is entirely enclosed by a little solid body. Although there is no pattern here that demands action, it is worth keeping an eye on. It signals indecision and whether or not the days that follow exhibit an upward or negative trend will determine what should be done.

   Bullish Harami

   The Bearish Harami's inverse indicates a downward tendency. The solid genuine body from the day before dwarfs the little open body. This indicates a break in a trend and it will take more time to spot any new patterns.

5. Bearish Harami Cross

   Open candlesticks in an upward trend are followed by a doji. This demonstrates that it is best to monitor the stock and look for any trends.

6. Bullish Harami Cross

   Solid candlesticks form a downward trend, which is then followed by a doji. This indicates a trend reversal and further information may be required before taking any actions, purchasing, or selling as a result.

4.4.2 Kagi Charts

An instrument for monitoring share and stock price changes is a Kagi chart. It advises stock traders on the best times to buy and sell shares.

Kagi charts, which were created in Japan in the late 1870s, were initially used to monitor rice prices so that dealers could purchase at the lowest cost. Steve Nison, who popularised Japanese candlestick charts to the West, also popularised the Kagi Chart.

He realised that Kagi charts gave advantages in comprehension and stock price movement analysis, giving traders a different approach to read and comprehend the trading matrix.

Kagi, which translates to "key" in Japanese, can be a useful tool for spotting shifts in emotion.

How Do Kagi Charts Work?

Kagi charts have several features that need to be understood before they can be deciphered correctly.

- The Z and Y axes are present. Dates on the horizontal X axis serve as indicators for significant price movements. The value scale is on the vertical Y axis.
- A thicker green line and a thinner red line are the two main sorts of lines. The yang line refers to the thicker green line. In essence, this is a "bullish" tendency towards the upside and an increase in demand over supply for the share. Increasing supply over demand is indicated by the thin red line, which is a yin line where the price drops below a prior waist. A "bearish" declining price trend is present.
- A waist is a horizontal line that connects a rising and a descending line.
- A shoulder is a horizontal line that connects a descending and ascending line.
- The line won't reverse until the share price moves by more than the pre-determined reversal amount, which is typically 4%. This reversal amount must be just big enough to demonstrate accurate price changes without being too big to miss signals.

Notes

- Depending on what the trader wants to see, this proportion can be extremely subjective. Instead of a percentage, a trader could set this pre-set number as a dollar amount or average true rating (ATR).
- When exceeded, the colouring of the shoulders and waists changes. When the shoulder or waist is passed, the hue changes. When there is a reversal, the colour does not always change; rather, it only does so when the cost reaches the waist or shoulder. A trader can rapidly determine if they should be buying, selling, or watching the stock by looking at a Kagi chart.

How Are Kagi Charts Used?

Kagi charts are easy to interpret. The thickness of the lines and the direction the line is moving in are the two things' traders should pay the most attention to. This line eliminates all the noise and irrelevant data because it doesn't change until there is a big price reversal.

A line travelling straight down on a chart won't change into a shoulder until a reversal amount of 4 percent (or whatever the trader has chosen) is reached. When the change was done and where it would finally continue to rise are then shown by the shoulder.

This will then turn green, signifying a strong rising price trend. A waist will form when the 4 percent reversal amount is attained and a downward price trend will then follow.

Traditionally, Kagi charts have been used to buy when the lines transition from thin to thick. There is no noise, thus it only displays trends, patterns and results. But because it's so basic, you won't be familiar with all the important aspects.

4.4.3 Open-High-Low Charts

An OHLC chart is a kind of bar graph that displays the opening, closing, high and low prices for each period. The four primary data points during a period are displayed on OHLC charts, which are helpful because many traders believe the closing price to be the most significant.

The chart type is helpful since it can display momentum that is either increasing or diminishing. When the open and close are widely apart, momentum is strong; when they are close together, momentum is weak or indicative of indecision. When determining volatility, the high and low show the entire price range for the time. On OHLC charts, traders look for several patterns.

A vertical line and two brief horizontal lines that extend to the left and right of the horizontal line make up OHLC charts. The initial price for the period is indicated by the horizontal line extending to the left and the closing price is indicated by the horizontal line extending to the right. The vertical line's height shows the intraday range for the time, with the high representing the high for the period and the low representing the low for the period. The entire thing is referred to as a pricing bar.

Since the close is higher than the open, when the price increases over time, the right line will be above the left. These bars frequently have a dark tone. The right line will be below the left line if the price drops during a period since the close is lower than the open. Typically, these bars are red.

Any time window can be used with OHLC charts. It will display the open, high, low and close price for each 5-minute interval when applied to a 5-minute chart. When used on a daily chart, it will display the daily open, high, low and close prices.

Notes

Compared to line charts, which just display closing prices connected into a continuous line, HLC charts display more information. The quantity of information displayed on OHLC and candlestick charts is identical, but they do so in somewhat different ways. Candlestick charts display the open and close using a real body, whereas OHLC charts use left and right facing horizontal lines.

Interpreting OHLC Charts

Technical analysts can analyse OHLC charts using several different methods. Here are some recommendations.

Vertical Height: The volatility during the time is indicated by the vertical height of an OHLC bar. If the line height is high, traders will be aware of the market's high volatility and uncertainty.

Horizontal Line Position: Technical traders can determine where an asset opened and closed in relation to its high and low by looking at the placement of the left and right horizontal lines. Traders can believe that the rally faded towards the conclusion of the session if the security climbed higher but the close was significantly lower than the high. If the price dropped but ended much above its low, selling dwindled as the time came to an end. The price couldn't move much either way, so if the open and close are close together, there is hesitation. If the closing is significantly higher or lower than the open, this indicates that there was significant buying or selling during the period.

Bar Colour: Typically, more black bars than red bars will be present during an uptrend. More red bars than black bars are typical during a decline. This can reveal details about the trend's intensity and direction. Briefly, there is strong upward movement shown by a series of broad black bars. This information may be useful when deciding whether to conduct additional analysis, even though it is still essential.

Patterns: Additionally, traders keep an eye out for trends on the OHLC chart. The key reversal, inner bar and outside bar are among the prominent patterns. When the price starts above the previous bar's close, hits a new high and then closes below the previous bar's low, a crucial reversal in an upward trend occurs.

It displays a major change in momentum that might signal the start of a retreat. A major reversal in a downtrend occurs when the price opens below the close of the preceding bar, makes a new low and then closes above the high of the previous bar. This denotes a material upward shift and an approaching rally.

Example of an OHLC Chart

This OHLC chart for the SPY represents the S&P 500 SPDR ETF. A greater number of black bars, like the time at the beginning of October, are often used to identify overall increases. By adding more alternate bar colours, the price rises marginally higher until mid-November but primarily sideways.

The price starts to increase in mid-November, which is shown by a couple of ranging black bars. Black rising bars predominated when the price increased at the beginning of the year. There are substantial red bands in the beginning of February, far larger than any of the earlier progress. This is a clear indicator of intense selling pressure.

4.4.4 Point and Figure Charts

In the point-and-figure approach of charting, the only variables are price changes and the direction of change. It has been around for more than 130 years and technical

Notes

analysts and traders frequently use it to predict future price movements. Charles Dow, The Wall Street Journal's founding editor, was one of the method's pioneers. However, as computer use became more widely available to regular traders, it has not been utilised frequently.

But the approach gives traders a roadmap that shows the conflict between supply and demand based on pure price movement without taking volume into account or taking time into account.

Dimensions of volume and time are not included. On graph paper, price information is plotted. When a reversal by at least one unit can be documented, new columns are started.

How to use point and figure charts

As we've previously explained, a point-and-figure chart essentially shows the price volatility of a stock over a selected time. Only units of price are displayed along the arithmetic y-axis. It shows how many times stock prices increased or decreased by a specific amount and the x-axis represents time periods.

Box size is the amount that was used as requested. It is directly connected to the distinction between the y-axis markers.

The only markings on the chart are stacked Xs and Os, each of which denotes a specific amount of price movement. Xs show how many times the stock increased by the predetermined limit, while Os show how many times it decreased by it.

If the price drops by a whole price unit (such as $0.50), a "0" is drawn. Once the price starts to move upward and changes direction, a "X" is placed in each box.

This filters out smaller price changes and allows traders to concentrate on trend quality.

What do point and figure charts tell you?

Based on previous price movement, point-and-figure chart patterns predict impending gains or losses. The optimal timing to purchase or sell financial products like stocks, options, ETFs and more can be determined by traders with the use of point-and-figure charts.

As was already established, the lack of a timeline along the bottom horizontal axis makes these infographics quite odd. They consist only of price fluctuations.

The fundamental idea behind point and figure charts is to only show the price when it makes a significant move. Nothing is charted if nothing notable occurs at a particular moment.

Because chart events like a reversal or breakout frequently coincide with actual events like breaking news, point and figure charts are event driven.

Pros of point and figure charts

1. Breakouts and breakdowns are clear cut: Signals to buy and sell are beyond discussion. However, a signal needs to be backed up by data, which can be seen in the chart.
2. Trader emotion contained: Trading decisions are made only based on mechanical point-and-figure signals, not on emotions or feelings. Similarly, if no signal is supplied, a morning news report won't sway a trader into acting.

Notes

3. No arbitrary drawing of trendlines: In contrast to other charting techniques like bar and candle, where trendline placing is arbitrary, point and figure charts' mechanical criteria for producing trendlines maintain consistency.

Point and figure charts also show significant and clear bands of support and resistance since noise is filtered out.

Cons of point and figure charts

1. P&F charts do not show gaps: As a result, nocturnal gaps are not apparent. For swing traders or intraday traders who use intraday charts, this poses a significant issue.
2. Signal generation is dependent on reversal method and box size scale: This implies that whether a signal prints or not depends on changes made to the chart's parameters. Therefore, it is crucial to use the same efficient box reversal mechanism and scale throughout all charts to maintain consistency between point and figure charts.
3. No volume on point and figure charts. Both traders and technical analysts respect volume. However, these charts omit volume, so you must look elsewhere, possibly from other sources, when you need more proof.

4.5 Types of Visual: Relationship

Visual relationships (such as "man riding bicycle" and "man pushing bicycle") represent a wide range of interactions between pairs of objects in photographs. As a result, the set of potential associations is very extensive and it is challenging to find enough training instances for all potential interactions.

This restriction has caused earlier research on visual relationship identification to focus mostly on predicting a small number of relationships. Even while most partnerships are rare, their objects (such "man" and "bicycle" or "riding" and "pushing") separately happen more often.

With this knowledge, suggest a model that trains visual representations of objects and predicates separately before combining them to predict multiple associations per image. Enhance previous work by adjusting the probability of a predicted relationship using linguistic priors from semantic word embeddings. From a few instances, our approach can scale to predict thousands of different sorts of associations.

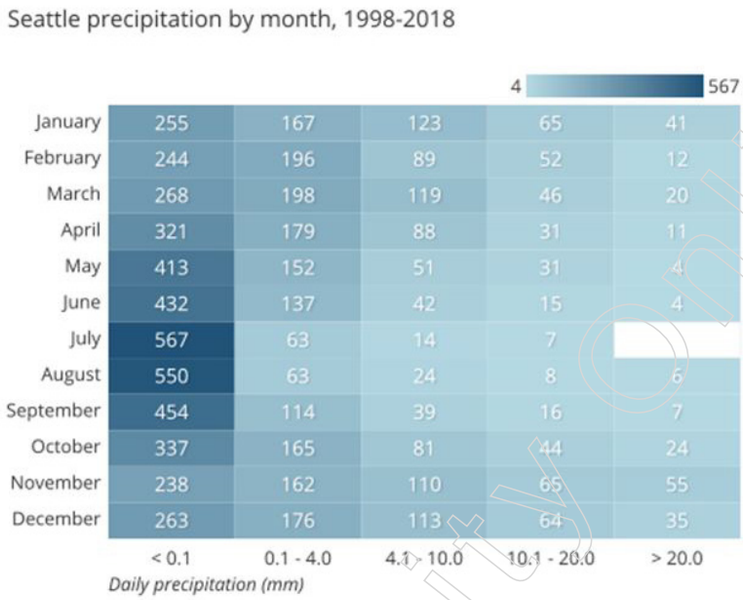
Localise the expected relationships' objects as bounding boxes in the image as well. Additional evidence that recognising relationships can enhance content-based image retrieval.

4.5. Heat Map

A grid of coloured squares, often known as a heatmap or heat map, shows the values for a significant primary variable over two axis variables. Similar to a bar chart or histogram, the axis variables are broken down into ranges and the colour of each cell reflects the value of the primary variable in the corresponding cell range.

The sample heatmap shown above shows the daily precipitation distribution for Seattle, Washington over an 11-year period, broken down by month. Like in a typical data table, each cell displays a numeric count. However, the count is also accompanied by a colour, with higher counts corresponding to deeper colourings.

Notes



Sample of heatmap

The deepest colourings in the left-most column of the heat map show that most days over the entire year had no precipitation. Additionally, the seasonal pattern of cell colours reveals that rain is more often in the winter months of November to March and less frequent in the summer months of July and August.

2-d density plots

When data is not restricted to a grid, the term “heatmap” is sometimes used in a broader sense. Website tracking tools, for instance, can be set up to observe how visitors interact with a page, such as by observing where visitors click or how far down, they typically scroll.

Notes

Every click (or other tracking event) has a place that radiates a tiny amount of numeric value in the vicinity of it. A comparable colourmap is displayed after all of these variables are totalled up across all events.

The output of these tools has a visual language that associates value with colour and is like the type of heatmap described at the top, except that it lacks a grid-based framework. Such heatmaps are also commonly referred to as 2-d density plots.

When you should use a heatmap

Heatmaps are used to show relationships between two variables using one variable on each axis. By observing how the colours of the cells vary across each axis, you can find any patterns in the values for one or both variables.

The grid cells for the displayed colours of the principal variable of interest in the latter case must be created by binning the numerical value, much like in a histogram.

Any number of metrics, including the frequency count of points in each bin or summary statistics like the mean or median for a third variable, can be represented by a cell's colour. The creation of a heatmap might be conceptualised as a table or matrix with colour coding on top of the cells. Cells may also be coloured in some applications based on non-numeric values, such as generic qualitative levels of low, medium and high.

Different visualisation tools may accept data in a variety of ways when plotting it as a heatmap. Data can be provided in one key form in the same way that a table would naturally show it. The first column of the heatmap will contain values for one of its axes and the titles of the subsequent columns will match to the bins for the other axes. The heatmap itself will be encoded with the values in those columns.

The other typical arrangement for heatmap data is a three-column setup. One row in the data table corresponds to each cell in the heatmap. The heat map cell's 'coordinates' are specified in the first two columns and the cell's value is specified in the third column.

Best Practices for using a Heatmap

1. Choose an appropriate colour palette

   Picking a colour scheme that goes well with the data is crucial because this type of chart relies significantly on colour. The most frequent association between value and colour is a sequential colour ramp, in which lighter hues indicate lower values and darker hues indicate higher values, or vice versa. The use of a diverging colour scheme is possible when values have a meaningful zero point, though.

2. Include a legend

   As a related point, a heatmap must typically contain a legend describing how the colours correspond to numerical values. A key is essential for viewers to understand the values in a heatmap because colour on its own has no inherent link with value. When the absolute relationship between value and colour is not significant, only the relative patterns of the plotted data should be included instead of a legend.

3. Show values in cells

   In comparison to other encodings like location or length, mapping colour to value is less precise. As a double encoding of value, it is a good practise to include cell value annotations in the heatmap whenever available.

4. Sort levels by similarity or value

   It can be worthwhile to think about altering the order in which those axis variable levels

Notes

are presented when one or both axis variables in a plot are categorical in nature. If there is no inherent ordering in the categories, you may want to choose an order that will make it easiest for the reader to understand the patterns in the data. One popular method is to order categories from greatest to smallest based on their average cell value.

5. Select useful tick marks

   There are options for how bins are organised and how they are displayed in the chart for numeric axis variables. Keep tick marks on each bin like for a categorical axis variable if there are only a few bins. To prevent congestion when there are many bins, it is preferable to plot tick marks between groups of bins. It may be a good idea to experiment with different settings because the size and number of bins you should use will depend on the data's nature.

Common Heatmap Options

1. Clustered heatmap

   It is a popular modification to have the horizontal axis show measurements of various variables or metrics rather than levels or values of a single variable. The result is something approximating a conventional data table, where each row represents an observation and the columns represent the entity's value for each measured variable.

   Since the objective of this sort of graphic is to create associations between the data points and their attributes, this type of heatmap is also frequently referred to as a clustered or clustering heatmap. With the same goal for variables, you wish to determine how individuals differ or are like one another.

   Clustering is typically used as part of the process by analysis tools that create heatmaps of this type. This application can be seen in the biological sciences, for example, when comparing the patterns of gene expression among different people.

2. Correlogram

   A correlogram is a heatmap version that shows a list of the numerical variables in the dataset in place of each of the variables on the two axes. Each cell shows how the intersecting variables are related, such as through a linear correlation. Sometimes, more intricate depictions of relationships, such as scatter plots, take the place of these straightforward correlations.

   Correlograms frequently play an exploratory function in statistical modelling, assisting analysts in understanding the relationships between variables.

4.5.2 Radar Chart

Using three or more quantitative variables, multivariate data that is mapped onto an axis is shown on a radar chart. It has a central axis with at least three radiating spokes that resemble the web of a spider. On these spokes, the values for the data are mapped. It aims to swiftly draw attention to differences, similarities and outliers for that product, service, or other interesting thing.

The terms irregular polygon, polar chart, spider chart, web chart, star chart, cobweb chart and Kiviat diagram are also used to describe it. This series of charts is credited to German inventor Georg von Mayr. In 1877, he published the first radar chart.

Notes

Take your favourite brownie as an easy example of how radar maps may be used. A brownie is made up of a wide variety of components: In addition to the crust, moistness and density, other characteristics include chewiness, chocolatey-ness, the inclusion of nuts and other components like cranberries.

Use a radar chart when:

- There are multivariate observations
- There is an arbitrary number of variables
- You need to identify outliers
- You need to make comparisons across products or services
- Data sets are small or moderately sized

When creating a radar chart, there are a few best practices:

- Variables should be arranged in some meaningful order
- More than three series should be presented on their own radar charts
- Don't use too many variables or the chart risks becoming confusing
- If there are multiple data series, the filled-in colour should be transparent

4.5.3 Venn Diagrams

A Venn diagram is a type of diagram that uses circles to show the relationships between objects or small groups of objects. In contrast to circles that do not overlap, overlapping circles have some properties.

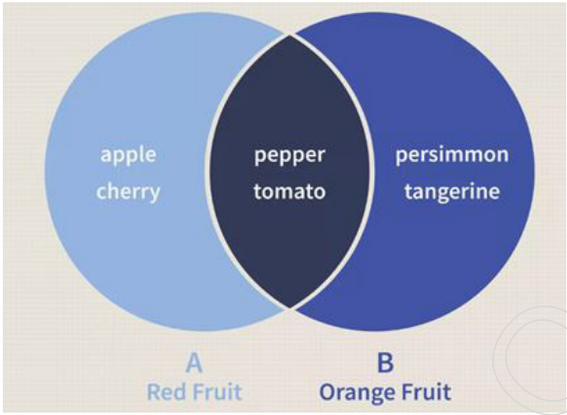
Venn diagrams are helpful for visually illustrating the relationship and differences between two concepts. Their worth as educational resources have long been acknowledged. Since the middle of the 20th century, Venn diagrams have been introduced into elementary school lesson plans and beginning logic curricula across the world. Venn diagrams are helpful for displaying the relationships between several ideas or factors. They can briefly illustrate how objects are alike or dissimilar as well as where and how they overlap. The intersection of two or more circles in the middle of a Venn diagram frequently depicts a nexus or central concept that may be broken down into the different other circles, with labels on the outer portions more general and distinct concepts than those towards the centre.

The overlapped regions can also be utilised to highlight the similarities between two seemingly unrelated settings. For instance, while the urban and rural contexts are unique with their own set of activities, you can see that they do share sporting events in the example figure below.

Examples of Venn Diagrams

A Venn diagram could be made to illustrate fruits that are red or orange in colour. You can see in the image below that some fruits, such tangerines and persimmons, are orange in colour (circle B), whereas apples and cherries are red (circle A). As shown by the area where the two circles overlap, peppers and tomatoes both have red and orange colouring.

Notes



Venn diagram

4.5.4 Arc Chart

Arc charts use arcs to depict links between source and target nodes (origin and destination) and represent nodes as circles on a single axis. A numerical column determines the arc's thickness, the colour designates the direction it is travelling in and the height designates the separation between nodes.

Source and target nodes, which are frequently used in the travel sector, display the starting point and ending point of the route taken.

When you create an arc chart, you specify:

- A source column
- A target column
- A numeric value column

If you are using a dive or marker as a source, it must have exactly two dimensions.

The nodes indicated in the source column can be sorted by the data source, by natural sort order (alphabetical or by numeric value), or both. The list of source nodes is followed by a list of target nodes that are not specified in the source column.

4.5.5 Chord Chart

A chord chart is a graphic way to show how data are related to one another radially around a circle. The data is represented as flows or connections between various entities (referred to as nodes), with the relationships between the nodes often depicted as arcs. In this case, the flow's significance is inversely correlated with the arc's size. Data from sophisticated scientific data to corporate use cases are visualised using chord charts.

Types of Chord Diagram

1. Bipartite Diagram

   Here, the nodes are divided into a few categories, but connections are only made between the categories.

2. Flow Diagram

   Both symmetric arcs per pair and two arcs per pair are viable representations for these maps. The dependence wheel is another name for this.

Notes

4.5.6 Tree Chart

A tree chart, also known as nesting, is a type of data visualisation that uses rectangles of progressively smaller sizes to show hierarchical data.

As suggested by their name, tree charts display data as a tree-like structure with readable branches and sub-branch levels. When presented in a compact, accessible and visually appealing manner, trees are good at taking a lot of raw data and enabling the user to rapidly discover trends and make comparisons.

Characteristics and Components of a Tree Chart

The following are the defining characteristics of tree charts:

- Rectangles are utilised to represent data.
- Two numerical values are represented by each rectangle. Some people refer to the rectangles as “nodes” or “branches.” The subsequent nested datasets are referred to as “leaves.”
- Rectangles’ sizes and plot colours are determined using the quantitative factors related to each individual rectangle.
- The data may be multi-layered. Hierarchically organised data is shown as a collection of nested rectangles, with the “parent elements” tiled alongside their “child elements.”
- The area size of the rectangle corresponds to the quantity when it is assigned to a category.
- The area of the parent category is equal to the total of its subcategories.
- The rectangles are put in the tree in order of size. The rectangles often have a size gradient from the top left corner to the bottom right corner of the chart. Because of this, the top left corner of the tree has the largest rectangle, compared to the bottom right corner of the chart’s bottom row.
- When dealing with nested rectangles, or hierarchical data, the same arrangement is used, with the lower-level rectangles placed inside of each higher level rectangle in the tree. The total of the nested rectangles’ areas determines the size and location of the parent rectangle that contains the nested rectangle on the chart.

A tree chart is made up of three main sections:

1. The Plot Area

   Each rectangle is coloured in tones representative to the top-level category, completing the visual representation’s body. However, other colour options may also be utilised when designing trees and this is only necessary when the tree solely displays data. The plot area is where visual representation takes place.

2. The Chart Title

   Your users will grasp the visualisation more readily if you give the chart a name that is both straightforward and descriptive.

3. The Legend

   The legend, which is commonly depicted by a sliding colour scale, is the area of the map that aids in differentiating one data series from another. Each colour represents one of the top-level categories (branches) in a legend with a colour key.

Notes

4.5.7 Network Chart

A network is a group of things connected by multiple nodes that act as placeholders for those objects. It is sometimes referred to as a network chart.

A network chart is a graphic depiction or diagram used in the technical sciences to show the relationships between various systems or devices that are represented by nodes in the network chart.

An edge or link, which refers to the connection of two items, joins these nodes together. The relationship between these nodes is depicted by lines connecting them or by edges. Network charts can be understood considering how the internet operates.

The internet can be visualised as a network chart, where various devices are linked together via wired or wireless connections to allow for the movement of the internet between them.

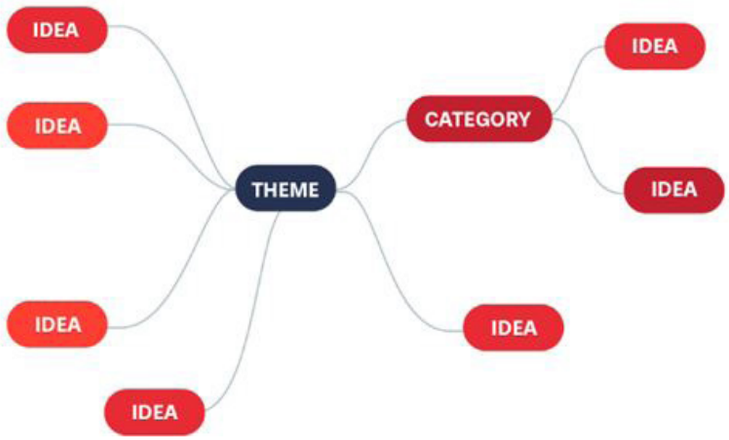
Because Network Charts are dependent on how a network operates, which incorporates many different components, it is crucial that comprehend how a Network Chart or just a network works.

Additionally, the edges and the flow of connections can be understood in terms of network charts. This idea of "Network edges" describes the information flow that can be divided into the following categories: -

1. Directed Edges: The different nodes or edges in a hierarchical flow are connected by directed edges, which have a specified and directed flow of connection.
2. Undirected Edges: Unorganised connections between undirected edges may flow from one edge to another in an unorganised manner.
3. Weighted Edges: The direction of a connection can also be decided by where an edge is placed. Perhaps some edges in a network chart are weighted because they have a higher value or weight than other edges.

4.5.8 Mind Map

A mind map is a visual organisation tool for information. The central idea of a mind map is typically a single idea. The standard format for presenting this idea is as a picture in the centre of a blank landscape page, to which related representations of the subject, such as images, words and word fragments, are added. Other ideas come from the major notions since they are closely tied to the fundamental idea.



Mind Map

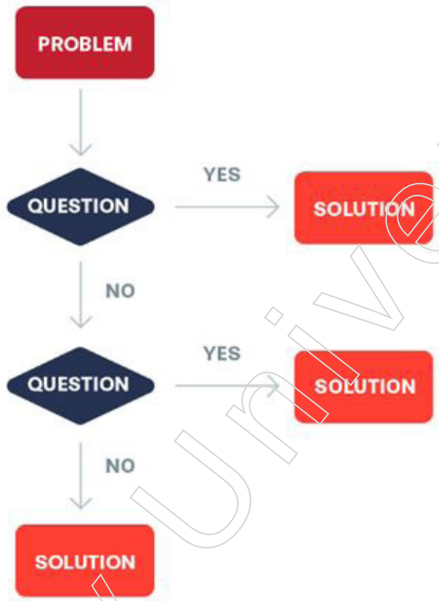
Notes

Making mind maps is a procedure used to communicate thoughts and concepts visually. Mind maps are nothing more than a "structured" visual depiction of your thought process. You could occasionally find yourself in a position where it is really challenging to persuade others of your point of view. Whether it be a manager, superior, or investor.

It's simple to make mind maps. Even a pen and piece of paper can be used to sketch it. Following is the general methodology: The main thought should be at the centre. Draw branches from the centre in a way that they relate to one another, showing the results at the very end.

4.5.9 Flow Chart

A flowchart is a sort of chart that illustrates a workflow or procedure. It shows the stages as different kinds of boxes and their order by connecting them with arrows. This diagrammatic image depicts a possible resolution to a particular problem. In many fields, flowcharts are used in the analysis, design, documentation, or management of a process or programme.



View of flowchat

This kind of diagram is used to display a process's consecutive steps. Using a network of interconnected symbols, flow charts depict a process, making it simple to grasp and easier to explain to others. A complex and/or abstract operation, system, concept, or algorithm can be explained using flowcharts. Making a flow chart can be useful for planning, creating, or refining a process.

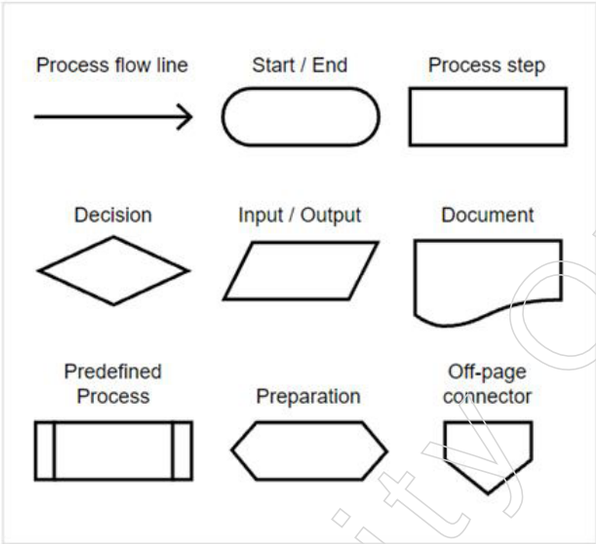
Symbols are categorised and standardised into various categories, each of which has a distinct shape. Inside the symbol form are labels for each step. A curved rectangle is used to represent the start and finish of the process in flow charts.

To depict the direction of flow from one phase in the process to another, lines or arrows are utilised. The symbol for straightforward instructions or tasks is a rectangle. While a diamond shape is employed when deciding. Other symbols that can be used in a flow chart include numerous others.

Flow Charts can run horizontally or vertically.

Notes

Flow Chart Symbols



Shapes of flowchart

Benefits of Flow Chart

The flowchart diagram is primarily utilised in a variety of problem-solving strategies because it presents solutions in the simplest, most easily understood and most easily recalled way possible. A flow chart has a variety of advantages. These are:

- Clarity in Representation: The use of flowcharts in documents allows for a clear visual representation of the progression of a program's events. The users are more aware of how to identify the necessary actions and remove the rest.
- Effective Communication: Flowcharts provide a step-by-step visual knowledge of every flow, assisting in successful global communication.
- Coordination: The efficiency of a flowchart contributes to the reduction of the overall burden of additional team members, including the capacity to plan and make events.
- Increase in Efficiency: The capacity of the flowchart to remove mistakes and pointless steps from a process aid in significantly improving each step of the process.
- Analysing: A flowchart aids in better problem analysis by illustrating the types of actions needed for each phase in a process.
- Solving Problems: A flowchart assists in breaking down a difficult problem into manageable, easily specified sections.
- Clarity in Documentation: The use of digital or programme flowcharts as paperless documentation improves the situation.

4.5.10 Waterfall Chart

A waterfall chart is a method of displaying data that demonstrates the potential impact of successive positive and negative values on an original value. Both sequential and category data can be displayed with this chart. It employs a succession of bars that display gains and losses to clearly demonstrate how an initial figure was altered by circumstances and resulted in the final value.

Notes

Due to its resemblance to bricks dangling in midair like the well-known video game Mario, it has also been referred to as a flying bricks chart, a bridge chart and a Mario chart. Nevertheless, even though water does not flow upwards, it is frequently referred to as a waterfall chart.

This chart's creation has been credited to McKinsey and Company, who first used it in presentations. Financial institutions frequently use this contemporary chart to display financial gains and losses.

Benefits of Waterfall Charts

1. Shows Changes
2. Simple to Understand
3. Gives Micro Stories

Disadvantages of Waterfall Charts

1. Hard to Compare Without a Baseline
2. Not Commonly Known

4.6 Types of Visuals: Proportion

Visualisation methods that use size or area to show differences or similarities between values or for parts to a whole.

4.6.1 Bubble Chart

When data needs a third dimension, bubble charts—also known as bubble plots or bubble graphs—are used to provide viewers with more information. To compare three separate variables, a bubble plot relational chart is utilised.

While a typical line graph might display the amount spent on a particular category of items (for instance, the dollar amount of sales of electronics), bubble graphs provide more details. The following example shows the value of furniture and electronics sales, with circles denoting the quantity of goods purchased.

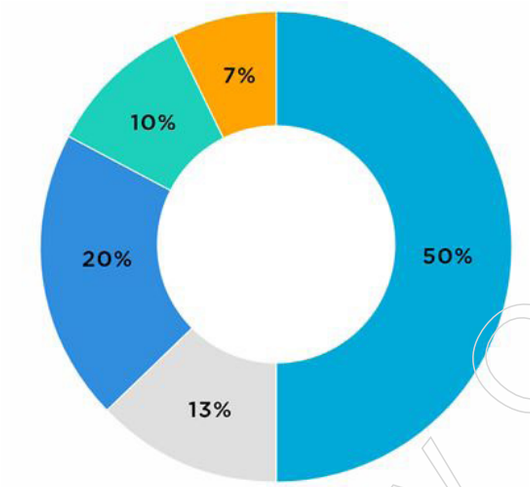
While bubble charts are useful for analysing relationships, they are not ideal for displaying precise data. Although the magnitude of a bubble's growth rate by itself cannot accurately reflect the rate of quantity expansion, it can provide readers an indication and set the stage for interpreting the growth rate in relation to the other two quantities.

The best use of a bubble chart is to answer a binary question, such as whether three variables are related or not. This connection may reveal a pattern.

4.6.2 Donut Chart

A donut chart is nothing more than a pie chart with the central section cut off to resemble a donut. At first glance, this can just seem like an aesthetic variance. However, a donut chart assists people in avoiding the confusion that a pie chart typically results in regarding the area parameter.

Notes



Donut chart

It is simple to mistake the area of each slide in a pie chart for the entire pie and draw conclusions based on this visual cue. Instead, the centre is absent in a donut chart, which encourages the reader to concentrate on the length of the arc rather than comparing it to the overall area that a circle would depict.

A donut chart also has the aesthetic advantage that the information inside the doughnut can be utilised to display data, labels and other things to make the chart easier to read. Different data bits are represented by each slice of a donut chart, which is frequently colour-coded for clarity.

Benefits of Using a Donut Chart

One of the simplest and most well-known ways to depict data is a donut chart. A donut chart is frequently the ideal choice when presenting information to vast and diverse audiences, provided that the data being represented is a whole set with numerous distinct pieces inside it.

A donut chart can be used in sales reports to evaluate the number of opportunities that are open, lost, or won and the revenue can be shown as a donut chart as a result. This enables decision-makers to determine whether leads that are won appropriately contribute to the bottom line or whether leads that are lost are too expensive to lose.

A dynamic donut chart can enhance this capability. When conducting forecasting, it can be helpful to know that a donut chart's appearance can alter depending on the values of the input data.

Donut charts can be created in various sizes, colours and arrangements, stacked on top of one another and named inside the chart area to conserve space. This makes the charts more interactive and provides users with access to more detailed data.

One of the most orderly methods of data representation are donut charts. A donut chart can be highlighted by making certain parts thicker to signify segments that are more significant. When using a donut chart, certain software programmes also automatically determine segment percentages.

Simply put, folks who are unfamiliar with dashboards and reporting may utilise a donut chart just as quickly and easily as their more experienced peers while still conveying the necessary insights.

Notes

Challenges When Using a Donut Chart

A donut chart can get crowded with too many segments, just like most other types of data visualisation. The segments may be difficult to read if there are too many to display and each one takes up a little percentage of the entire data.

The ideal format to use when representing negative values is not a donut chart. For instance, debt can only be shown as a part of the entire and not in terms of how it affects overall cash flow in a financial planning donut chart.

Additionally, even though a donut chart is excellent for comparing data, analysis with just a donut chart is frequently challenging because there are no other ways to interpret the chart except through visual clues. But by adding the percentage labels adjacent to each chart section, this problem can be resolved.

A donut chart is not the greatest choice to use if changes over time need to be recorded since, unlike a bubble chart's tracker, it cannot keep the information it depicts over time.

4.6.3 Marimekko Chart

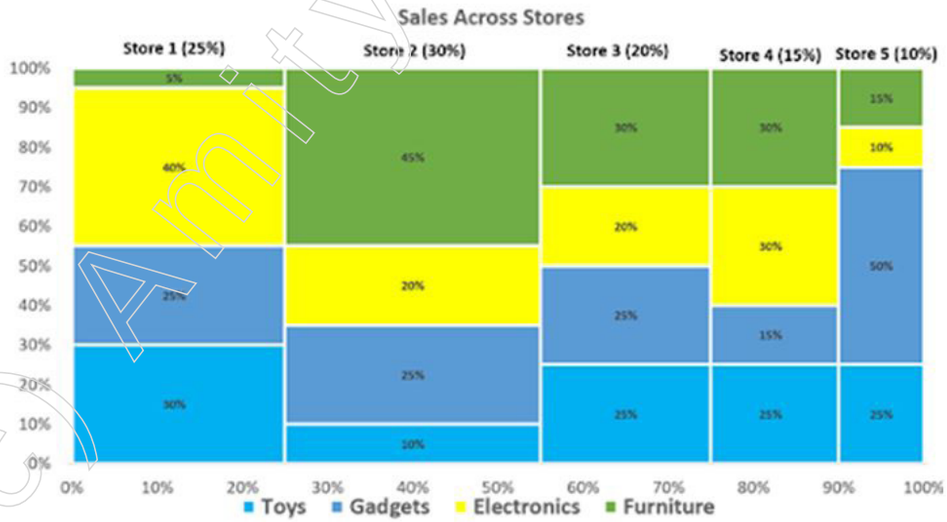
A graphic depiction known as a Marimekko chart makes use of stacked bar graphs of various widths to show category data. The mosaic plot or simply Mekko charts are other names for Marimekko charts. They are perfect for categorising sample data.

When Should Marimekko Charts be Used?

1. To Represent Sales Across Multiple Stores

   Marimekko charts are a useful tool for aiding in the visualisation of sales data. Assume a business owns a sizable chain of retail establishments spread out over several different regions. These shops provide products from a variety of categories, including toys, accessories, technology and furnishings. The categorization information of retailers and the items for sale can be represented by a Marimekko chart.

   For example, the graph below displays sales information from several stores. On the vertical axis, the total sales are distributed among the various stores. A horizontal axis with many classifications is shown. This graph makes it clear which retailer had the best sales and which product had the highest sales to volume ratio.



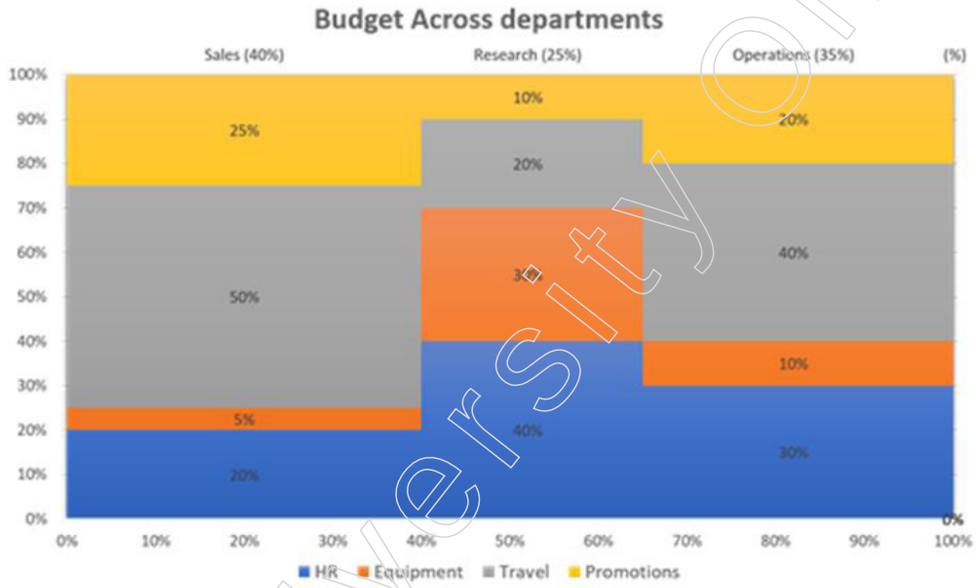
Marimekko Chart

Notes

2. To Compare Budget Breakdowns

The financial allocations made by various departments within an organisation can be visualised using marimekko charts. The vertical axis in the following graph denotes departments: sales, operations and research.

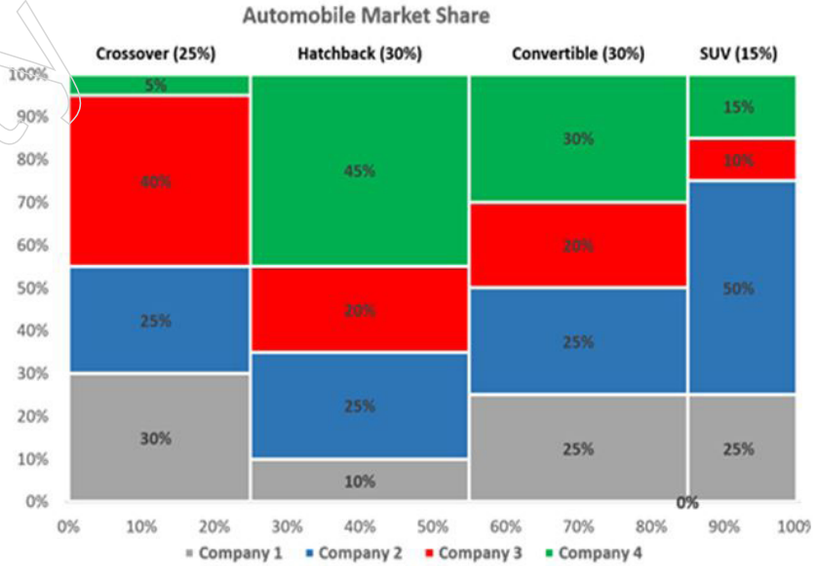
The human resources, equipment, travel and promotions budget allocations for each department are shown on the horizontal axis. The chart below shows how each department's budget differs depending on its goals and functions.



Budget in Marimekko Chart

3. To Represent Market Share Across Categories

The market share of various enterprises in a market with many segments is frequently represented using marimekko charts. The following graph displays the market share held by major automakers across various car segments: Hatchback, convertible, coupe, crossover and SUV.



Automobile Marimekko chart

Notes

4. To Represent a Skill Matrix

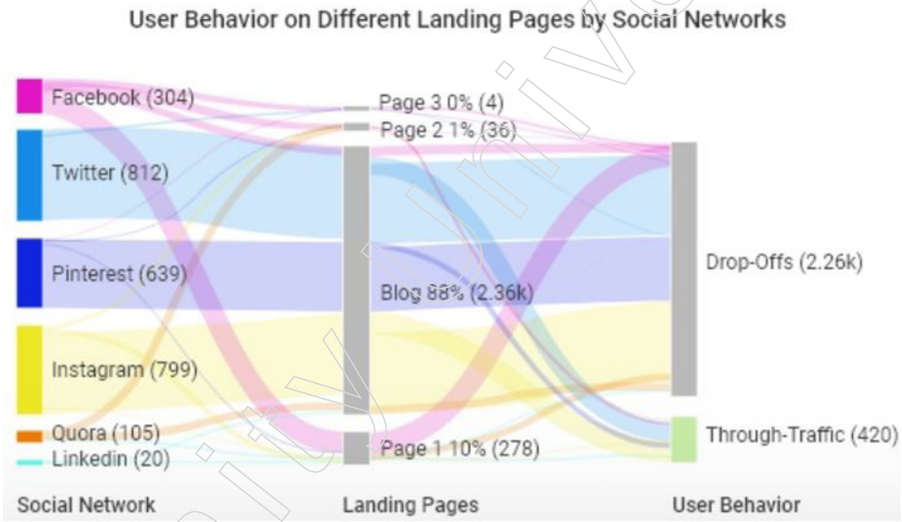
A wonderful tool for illustrating the skill set and level of competence required for various job roles is a marimekko chart. In a big data organisation, for instance, a Marimekko chart can represent important job responsibilities and their associated abilities, such as data scientist, statistician, developer and researcher.

4.6.4 Sankey Chart

You may get a thorough, high-level understanding of how your data moves and changes from one stage to the next with the Sankey Chart.

- Monitoring these motions reveals a wealth of important information, including
- The areas where you spend the most cash or resources
- How tight or lose your funnel is
- The most significant changes between stages
- How data flows and changes from start to finish of a process
- And more

Sankey Charts have numerous uses and can save you time while performing visual analyses and making decisions. The direction and amount of data via different phases, categories, or stages are displayed on a Sankey Chart, also known as an energy flow chart. The Sankey Chart has a wide range of uses and functions today, having originally been developed to literally visualise how energy flows in an engineering system.



Sankey Chart

What Is a Node in Sankey Charts?

Sankey Diagrams depict the movement of energy through one or more “nodes.” The category or stage where resources or energy go to, from, or through is known as a node.

Nodes are represented by vertical bars at each level in a standard Sankey Chart design. Each node’s length corresponds to the component’s value or size. Imagine a situation in which you would utilise a Sankey Chart to assist illustrate what a node is.

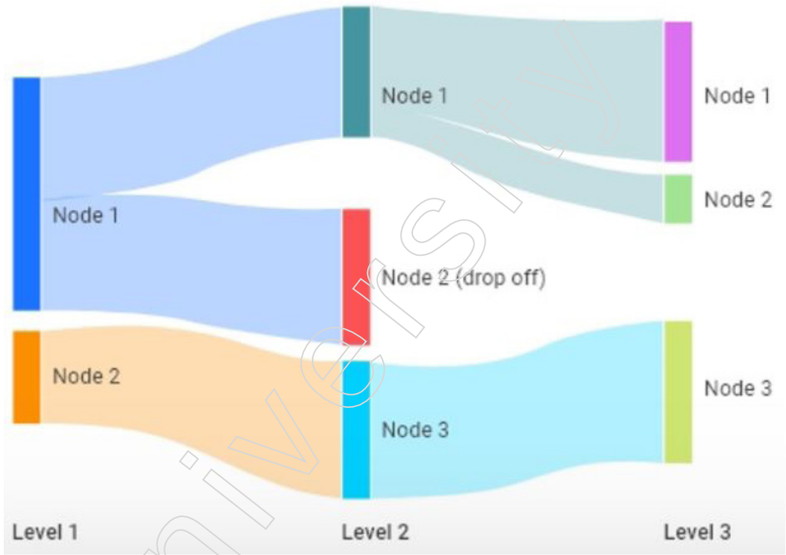
A marketer is interested in knowing the gender, age and device characteristics of the people who click on their adverts. With these data, a Sankey Chart would require three

Notes

levels with several nodes in each level. Your device kinds would be the first level. You might have three nodes for desktop computers, mobile smartphones and tablets here.

The gender level may then have 2 additional nodes for males and females. Finally, you would have the necessary number of nodes to account for each age range. Once more, the size of the node indicates how important that item is; a larger node is considerably more important. You might infer from this marketing scenario that the mobile node would be the largest, followed by the desktop and tablet nodes.

Understanding levels is also essential. Each level could stand for several data categories, stages of a system or process, a certain period, etc. However, when you include more nodes and levels, your straightforward Sankey Chart becomes more intricate.



Nodes and level in Sankey Chart

What Does Each Arrow Represent in a Sankey Visualisation?

Arrows, usually referred to as connections, are situated between nodes and levels. These arrows depict the data's energy flow. Always keep in mind that "energy" can refer to resources, materials, or other measures. You may increase the efficacy of your results by paying attention to the size and path of each arrow, which provide a wealth of useful information.

Link size: This conveys how much vitality or value this thing has. Your most important elements are represented by the Sankey graph's most noticeable linkages or arrows.

Link direction: The trajectory or course of each link will alter depending on how your data travels through various nodes. Links may also break if data is being sent through several categories.

Link colour: Each link's colouring serves two functions. It first facilitates the separation of each arrow from the others. When numerous linkages overlap, this is essential. You can also tell which node the link is originating from by looking at its hue.

Many linkages can produce a complicated Sankey chart, much like nodes and levels do.

4.7 Case Study

Notes

1. Data Science in Healthcare

The development of AI has had a significant positive impact on the healthcare industry. Healthcare practitioners have been using data science, particularly in medical imaging, to help patients receive better diagnosis and treatments. Several cutting-edge healthcare analytics technologies have also been created to produce clinical insights for enhancing patient care.

Additionally, by assisting in the creation of medicines tailored to individual patients, these instruments help reduce clinics' and hospitals' operational costs. In the healthcare sector, Natural Language Processing (NLP) is widely used to assess published textual research data instead of medical imaging or computer vision.

BioTech

How AstraZeneca harnesses data for innovation in medicine

AstraZeneca is a well-known biotech business that uses data analytics and artificial intelligence to find and deliver new, more potent drugs more quickly. To effectively treat diseases including cancer, lung disease, heart, renal and metabolic ailments, their research and development teams are employing AI to analyse huge data.

They can find new targets for cutting-edge drugs using data science. They decided to work with BenevolentAI on the first two AI-generated therapeutic targets for chronic kidney disease and idiopathic pulmonary fibrosis in 2021.

AstraZeneca is also using data science to improve clinical trials, establish personalised treatment plans and innovate the process of creating new medications. By 2026, its Centre for Genomics Research will evaluate almost two million genomes using data science and AI.

In addition, they are teaching their AI systems to scan these photos for disease and biomarkers of potent treatments. They can analyse materials more correctly and easily with this method. Additionally, it can reduce analysis time by about 30%.

By examining the clinical trial data, AstraZeneca also uses AI and machine learning to optimise the process at various stages and reduce the overall time for the clinical trials. In conclusion, they employ data science to improve patient care and drug development strategies, create innovative medicines, create smarter clinical trial designs and many other things.

2. Data Science in Hospitality Industry

Travel industry and data science

The travel sector benefits from predictive analytics in many ways. These businesses can increase personalisation and enhance customer interactions by combining recommendation engines with data science. To boost sales and boost revenue, they can research products and cross-sell them by suggesting related products.

Additionally, data science is used to analyse social media posts for sentiment, providing priceless travel-related information.

These agencies can better understand user demographics, the expected experiences of their target audiences and other things by knowing if these views are good, negative, or neutral. These data are crucial for creating aggressive pricing

Notes

strategies that attract clients and allow for greater consumer customisation of travel packages and related services.

Predictive analytics is used by online travel firms like Expedia and Booking.com to provide customised recommendations, new items and efficient product marketing. The same strategy helps not just travel agencies but also airlines. Due to flight disruptions, delays and cancellations, airlines routinely suffer losses.

Data science enables them to recognise patterns and anticipate potential bottlenecks, successfully minimising losses and enhancing the entire travelling experience for customers.

Summary

- Software libraries or frameworks known as “data science packages” offer functions and tools to make data analysis, manipulation, visualisation and modelling easier. These tools are made to make it simpler for data scientists to use different analytical methods and work with data
- A lot of data scientists work using high-level programming languages. Those who want to work in the subject may want to start by specialising in a data science language
- The most popular programming language for data science nowadays is Python.
- A sophisticated programming language created by statisticians is called R.
- Almost every industry nowadays relies heavily on data science, whether it is for planning, future forecasting, or making business decisions,
- A package in Python is a grouping of modules. Modules that are connected to one another are typically packaged together.
- An API is a (fictitious) contract between two pieces of software that states that if the former accepts input in a certain format, the latter will increase its capabilities and deliver the results to the former.
- A bar chart, often known as a bar graph, is a graph that uses rectangular bars with heights corresponding to the values they represent to show categorical data.
- Kagi charts, which were created in Japan in the late 1870s, were initially used to monitor rice prices so that dealers could purchase at the lowest cost. Steve Nison, who popularised Japanese candlestick charts to the West, also popularised the Kagi Chart.
- A Venn diagram is a visual representation that makes use of circles to highlight the connections between different objects or limited groups of objects.

Glossary

- Web API: n interface to a web server or a web browser is referred to as a web API.
- SAS: SAS is the name of a tool that ought to be first in the “statistical” category while you’re working with visualisation.
- R: A sophisticated programming language created by statisticians is called R.
- Data Visualisation: The graphic display of information and data is known as data visualisation.
- Pie Chart: A pie chart is described as a circular chart with numerous sections, each of which represents the percentage that each value contributed to the overall value.

Notes

- Area chart: A graph that combines a line chart with a bar chart to depict changes in amounts over time is called an area chart
- Candlestick chart: A candlestick chart is a type of financial graph that often displays changes in the price of commodities, equities, or derivatives.
- Heat Map: The values for a primary variable of interest over two axis variables are shown as a grid of coloured squares in a heatmap.

Check your Understanding

1. _____ is currently the most widely used computer language for data science. It has been in use since 1991 and is a simple, open-source language.
   a) Python
   b) JavaScript
   c) C++
   d) R
2. Web programming and machine learning are two examples of ____ applications.
   a) R
   b) Scala
   c) Julia
   d) SQL
3. Which language is great for working with matrices.
   a) Python
   b) Julia
   c) C++
   d) SQL
4. Which is the priciest software in the market, only large-cap companies would require you to be familiar with this tool?
   a) Microsoft Power BI
   b) Tableau
   c) Knime
   d) SAS
5. Which tool is an open-source data analysis tool created in Java and based on Eclipse?
   a) Tableau
   b) TensorFlow
   c) Knime
   d) Snowflake
6. Which is the main Python tool for scientific computing?
   a) NuMPy
   b) SciPy
   c) TensorFlow
   d) Pandas

Notes

7. The Twitter REST API, Facebook Graph API, Amazon S3 REST API are the example of which API.
   a) Operating System
   b) Database system
   c) Hardware system
   d) Web based API
8. Which bar chart are arranged from highest to lowest incidence?
   a) Stacked bar chart
   b) Grouped bar chart
   c) Column bar chart
   d) Pareto Chart
9. A graph that combines a line chart with a bar chart to depict changes in amounts over time is called an ____.
   a) Bar Chart
   b) Area Chart
   c) Cluster Chart
   d) Scatter Chart
10. Which chart displays the correlation between two variables?
    a) Scatter chart
    b) Bar chart
    c) Whisker chart
    d) Area Chart
11. Which is/are the application/s of clustering technique in Machine Learning?
    a) Biology
    b) Land use
    c) Search engines
    d) All the above
12. In which type of chart shows how data flows through a process?
    a) Funnel chart
    b) Density chart
    c) Cluster chart
    d) None of these
13. When is Candlestick chart created by a Japanese guy named Homma?
    a) 1800
    b) 1900
    c) 1600
    d) 1700
14. Which chart is used for tracking price movements for shares and stocks?

Notes

a) Candlestick chart

b) Kagichart

c) Open high low chart

d) Radar Chart

15. A _____ is a heatmap version that shows a list of the numerical variables in the dataset in place of each of the variables on the two axes

a) Correlogram

b) Clustered heatmap

c) 2-d density plots

d) None of the above

16. What is/are the other term used to describe Radar chart?

a) Polar chart

b) Star chart

c) Kiviat diagram

d) All the above

17. While creating an arc chart what must be specified.

a) A source column

b) A target column

c) A numeric value column

d) All the above

18. A _____ chart is a graphic way to show how data are related to one another radially around a circle.

a) Tree Chart

b) Chord Chart

c) Radar Chart

d) Arc Chart

19. Which chart makes use of stacked bar graphs of various widths to show category data?

a) Marimekko Chart

b) Sankey Chart

c) Stacked Graph

d) Tree Chart

20. A Python 2D plotting toolkit called _____ makes it simple to create cross-platform graphs and charts.

a) SciPy

b) Pandas

c) Matplotlib

d) Keras

Notes

Exercise

1. Explain the concept of API.
2. What are the different types of visual comparison?
3. Describe the visual patten and its type
4. Explain the visual relationship.
5. What is the different type of visual proportion?

Learning Activities

1. Explain all the tools of data science with example.
2. Explain packages of data science with example.

Check your Understanding-Answers

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

Further Readings and Bibliography

1. Field Cady. The Data Science. 2017
2. William Vance. Data Science: 3 Book in 1 – Beginner's Guide to learn the Realm of Data Science. 2020
3. Peter Bruce Andrew Bruce. Practical Statistics for Data Scientist. 2020
4. Reema Thareja. Data Science and Machine Learning using Python. 2022
5. Uma Maheshwari R Sujatha. Introduction to Data Science: Practical Approach with R and Python. 2021

Module - V: IBM Watson Studio and Jupyter Notebook

Learning objectives

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

- Understand Jupyter notebook
- Analyse IBM Watson Studio
- Describe Python
- Discuss visualisation packages

Introduction

Data scientists, developers and analysts can create, operate and maintain AI models, as well as optimise choices from any location with IBM Watson® Studio. On an open multicloud architecture, bring teams together, automate AI lifecycles and shorten time to value.

Python is one of the more than 40 programming languages that Jupyter supports. Installing the Jupyter Notebook itself requires Python (Python 3.3 or higher, or Python 2.7).

5.1 Programming Packages

Simply said, a package is a means to ensure that none of the names you select to use in your programme “step on the toes” of names that another programme might use. This is especially problematic when using “libraries” (sets of pre-written code) in huge software systems.

In the end, packages are merely a shorthand for naming things. You could give all your classes’ names like “jims_star,” “jims_bank,” “jims_this,” and “jims_that,” but that would be laborious. Instead, we insert a package declaration (such as package Jim) at the very top of a file and the computer automatically renames everything for us.

In the context of software, a package is a module that may be added to any programme to offer new features, capabilities, or options. As shown in the following Java code, a package can frequently be added to a programme using a “include” or “import” sort of expression.

import java.io.FileReader;

The FileReader package is used for reading character streams, useful for obtaining user input from a console.

5.1.1 Selection of Software

Every aspect of your organisation should be impacted by an effective system. These in-demand business applications support the management of various tasks, such as planning, research and development, purchasing, supply chain management, sales and marketing.

Numerous options are accessible, so it’s imperative that you pick the one that works best for your business. But do you understand how the selection procedure works?

Notes

Selection Process

We'll help you through the software selection process to get things starting. This simple, four-step method aids you in identifying your company needs, assessing and contrasting potential solutions, validating your technical requirements and contract negotiation.

1. Requirements and Research

   Consider your set of specifications as a checklist for selection criteria. What your company requires from a software solution is called a need. You can use these factors to determine whether a system is the best match for your company.

   Requirements Gathering

   When evaluating corporate requirements, keep in mind that a software's main advantages are its broad data analysis, precise and detailed reporting tools and numerous process automation options.

   Preliminary Research

   You can now carry out exploratory study to ascertain which solution would work for you after having your precise business requirements in hand. Most of this research may be done online, but you can also consult vendors and ask your employees who are familiar with software for suggestions. If you want to find the right candidate for your firm, consider using consultants as a resource.

2. Vendor Comparison and Sourcing

   Informal Inquiries

   Use the knowledge you've gained from gathering requirements and conducting research to assist you ask pertinent questions of potential vendors.

   - ❖ Verify if the vendor's product is compatible with your current legacy systems.
   - ❖ To maximise the likelihood of a smooth transition, find out if the vendor has experience in your business.
   - ❖ Confirm that the vendor fully comprehends your company's requirements and give the vendor any project plans you may have created.

3. Technical Validation

   Once you've asked suppliers for information and gotten their responses, it's essential to thoroughly assess the systems that are available..

   Technical Evaluation Scorecards

   Comprehensive scorecards provide an in-depth analysis of every feature and capacity of the systems, properly and comprehensively evaluating them. Cost, meeting user needs, streamlining internal procedures and adaptability should be the minimum number of factors you utilise to evaluate the system.

   Demos, Proof of Concept and More

   After using tech evaluation scorecards and proposal specifics to narrow down your list of potential vendors, it's time to ask the vendors for demos. Along with demos, ask for documentation that explains how the system will satisfy your company needs.

4. Financial Due Diligence

Notes

Business Case Review

A system's capabilities and cost should be compared to your organisation's budget, your organisation's budget and the anticipated savings to ensure that it will benefit your business and is now financially viable. Consider how much a project will actually cost to implement.

Evaluating Potentially 'Hidden Expenses'

You should factor in the costs of updating software, changing hardware and performing routine maintenance before selecting a system.

Contract Negotiation and Close

It's time to negotiate your contract after you've selected your vendor, gotten your price, looked through your proposal and finished reviewing your case.

This tactic might be labour- and time-intensive, but it will be worthwhile if you carefully choose your programme.

However, we offer technical evaluation scorecards for extra careful documentation to make things a little easier. You'll soon discover the ideal answer for your company.

Selection Criteria

Now that you are familiar with how the selection procedure operates, let's discuss the goal of this essay.

The most important factors for choosing software are listed below. Throughout all phases of the selection procedure, you can use these criteria to decide whether a solution is suitable for you.

1. Functionality and Ease of Use

   Have the following questions ready to consider first: What features and how well-designed is the system? And what features and level of usability do you require?

   You can assess if the product is fundamentally the correct fit for your company by responding to these inquiries.

   This phase may be the most time-consuming and tedious to accomplish because it needs a thorough analysis of your company and the product but hang in there. This is the crucial task.

   You can better comprehend the programme, compile your needs, or contrast systems by doing the following things:

   - Requirements Gathering – Think about the issues your organisation is now facing and how a system might be able to resolve or decrease them. then consider whether system features will allow you to accomplish this.
   - Consider Business Process Automation – Consider how many of your company's typical business operations could be automated and assessed if the solution has the necessary tools.
   - Examine End-Users – Do your end users have the ability to employ sophisticated software features and how tech-savvy are they? Do they currently use a system? You can decide whether a product's user-friendliness meets your needs by responding to these questions.

Notes

- ❖ Plan for Centralization – Consider the systems you already use that will need to be integrated with your system and inquire about integration alternatives.

2. Vendor Viability

It's a good idea to think twice before investing in a certain solution and to consider the vendor's reputation and your ability to work with them over the long run. If a corporation isn't reliable and secure, a product should be off limits even if it has all the necessary functions.

By examining vendor viability, you can be confident that your business will have a great partner and a great system. Consider the following factors when contrasting systems and the businesses that sell them:

- ❖ Company Credibility
- ❖ Product Viability
- ❖ Scalability

3. Technology

Systems make extensive use of technology; for instance, standard software should contain functionality for business intelligence, reporting and customization. You should take into account the technologies the solution offers and those that are necessary, such as how functionality is assessed?

4. Cost

It might seem clear that you should take a system's price into account, but it's not always as easy as it seems. It's crucial to question yourself: With so many different factors to take into consideration: How much will the actual price of the solution be?

Consider the following factors while evaluating the cost of a solution:

- ❖ Basic Pricing Information – First, consider the platform's price and determine if it is acceptable considering the features and technologies it provides. Because there are many options accessible, you can usually choose another product if the vendor's pricing policy is unreasonable.
- ❖ True Cost – Determine the real cost of the solution next. Determine the long-term total cost of ownership (TCO), considering all relevant expenses.
- ❖ ROI – Estimate whether the anticipated return on investment (ROI) will be significantly higher than the TCO once the latter evaluations are complete.

5. Support and Training

The two components of the puzzle that always get a system up and running and working are support and training. There are several options available on the market today for assistance and training from suppliers.

Alternatives to in-person training, access to training videos or materials and a knowledgebase or user community online are frequently included in a comprehensive training plan. Frequently, comprehensive support includes access to a contact centre and many online resources.

On the other end of the scale, some vendors provide neither support nor training, or they outsource both.

Considering the following, it's crucial to look closely: What kind of assistance and instruction will be given? What kind of assistance and instruction are required? Review

Notes

the available support and training options, make sure your end users would benefit from them and create a clear contract with the vendor.

6. Industry Expertise

The ability of a vendor to meet your criteria is fundamentally related to their understanding of the industry. Depending on your demands, you might be able to find a solution that's specifically suited for the nature and size of your organisation.

It's beneficial to work with a provider who is familiar with how your business runs because it allows the system to incorporate things like industry best practises. In light of your company's nature, do you require a specific platform

Expertise – Think about how long the vendor has been active in your industry and whether they have a track record of success there.

7. Implementation

Launching your new system requires implementation. A good approach is to find out from the seller how they will ensure a successful deployment. This stage will be especially important if you anticipate a challenging implementation, such as when switching from an old system to a new one.

- Implementation Assistance – Estimate whether the anticipated return on investment (ROI) will be significantly higher than the TCO after you have completed the latter analyses?
- Implementation Partner – If the vendor doesn't offer implementation support, you should usually collaborate with an implementation partner.

Our list of factors for choosing software is now complete. This list can be used to compare options during the full software decision process.

The process of choosing a system is lengthy and involves several aspects. But as you can see, it can be divided into smaller, more manageable chunks. Follow the procedure we've detailed step by step and don't forget to start by gathering your requirements.

You'll choose a vendor after careful consideration that will supply the system needed to advance your company.

5.1.2 Jupyter Notebooks

The free and open-source Jupyter Notebook web application allows you to create and share documents with live code, equations, visuals and text. The Jupyter Notebook is maintained by Project Jupyter employees.

Documents for technical and data science material are called Jupyter notebooks. An overview of Jupyter notebooks, its components and how to use them is given in this tutorial.

Utilise Datacamp Workspace, a hosted notebook service that offers all the features of Jupyter notebooks as well as features for connecting to databases, real-time collaboration and publishing your work, to learn more about notebooks.

In a single document, notebooks integrate computer code (such as Python, SQL, or R), the results of the code's execution and rich text features (such as formatting, tables, figures, equations, links, etc.).

The ability to provide commentary with your code is the main advantage of using

Notes

notebooks. The error-prone practise of copying and pasting analysis results into a different report can therefore be avoided. Instead, you merely incorporate yourself into the report's text in the notebook.

Who should use Jupyter Notebooks?

Data professionals, notably data analysts and data scientists, use Jupyter Notebooks the most. Over 80% of respondents to the Kaggle Survey 2022 indicated that Jupyter Notebooks were their preferred data science IDE.

Types of Jupyter Notebook

Hosted and local notebooks are the two primary variations of Jupyter Notebook. We will utilise a hosted Jupyter Notebook from DataCamp Workspace for the most of this session. For professionals and students who don't want to set up a local environment, workspace is a great choice.

Components of a Notebook

Three essential parts make up a Jupyter Notebook: cells, a runtime environment and a file system.

The notebook's individual units are called cells and they can either contain text or computer code:

- Narrative material is written in text cells, along with equations, links and images.
- Text cells are written using the straightforward markup language Markdown.
- Code is written and executed using code cells.
- Directly beneath the code cell will be displayed the output from the code cells.
- SQL cells (Workspace only) are used to run SQL queries, allowing you to access database data with ease.
- Chart cells can be used to easily visualise Pandas dataframes and generate visualisations (Workspace only).

The code in the notebook is executed by the runtime environment. One can design the runtime environment to accommodate several languages, such as Python, R, or SQL.

You can upload, store and download data files, code files and analytic outputs using the filesystem.

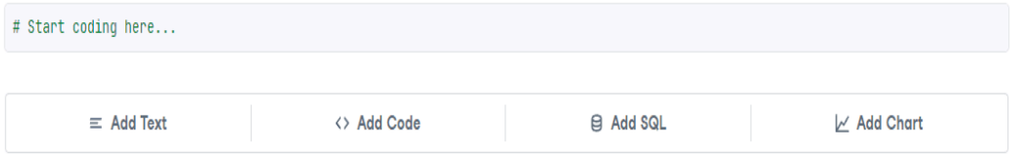
Command mode and edit mode

There are two main ways to interact with Jupyter notebooks: command mode and edit mode. You can move between cells, add and remove cells and modify the cell type when in command mode. You can modify a cell's content while it is in edit mode.

You can either click outside a cell or hit Escape to enter command mode. You can click inside a cell or hit Enter to activate edit mode.

In Using the 'Add Text' or 'Add Code' buttons in the workspace, you can add a new cell.

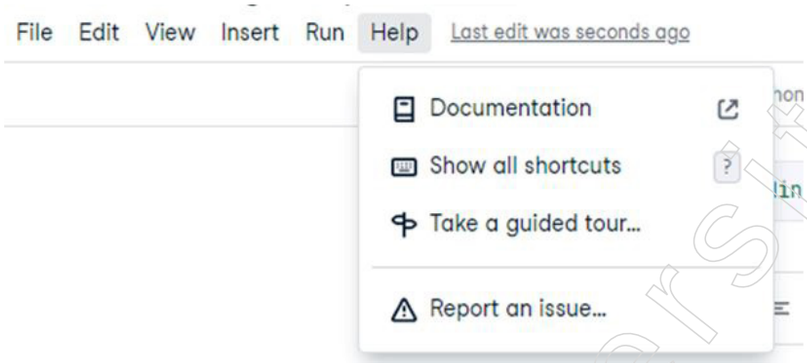
Notes



Jupyter Notebook console

Getting help

You can receive assistance for Jupyter notebook by using the menu option or the documentation. By selecting the help option from the menu, you may easily access help and keyboard shortcuts in Workspace.

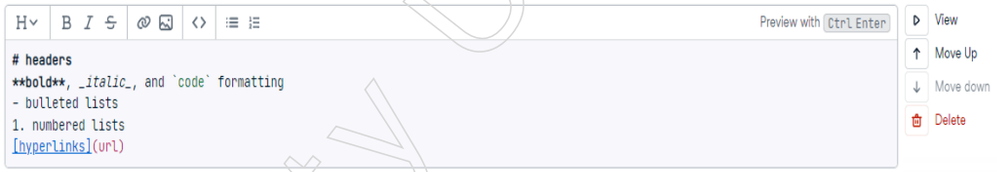


Jupyter Notebook console

Writing text

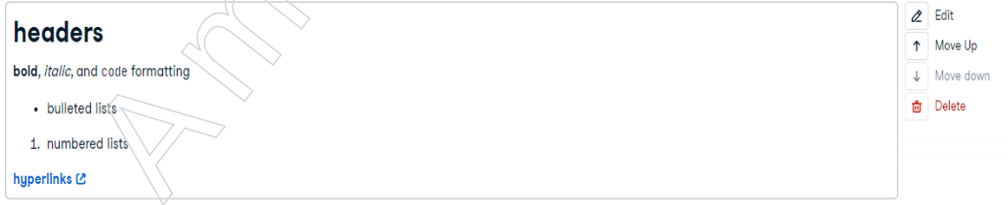
The Markdown markup language is used in text cells, making it simple to write and format content. When in edit mode, you can format your text using the buttons or syntax, such as ** ** for bold.

Here are a few different options:



Jupyter notebook console

Pressing shift + enter or the 'View' button will run the cell, giving the following result.



Jupyter notebook console

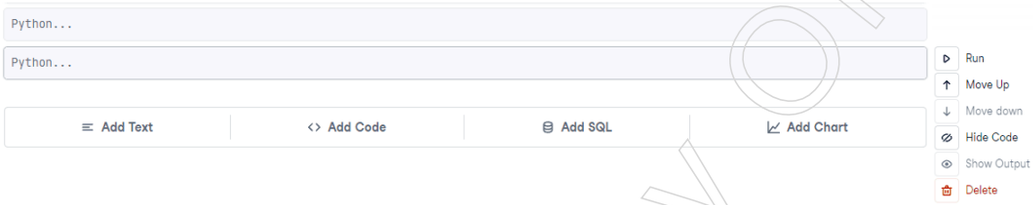
- Lines that start with a # are top-level headers. For a second-level header, start with ##, followed by ### for a third-level header and so forth.
- To make text bold, italic, or code-formatted, enclose it in **, __, or '.

Notes

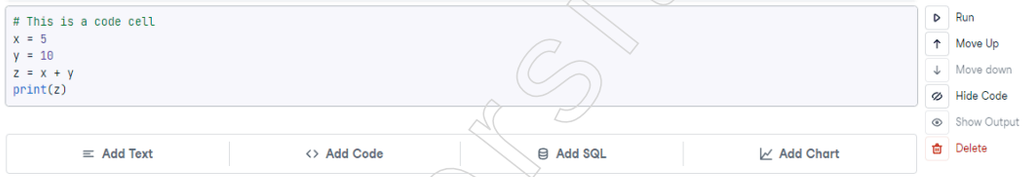
- To build a list of bulleted items, begin successive lines with a -.
- To turn a line into a numbered list, start it with a number, then add a period.
- Hyperlinks consist of two sentences. The URL is enclosed in brackets after the text to show is enclosed in square brackets.

Writing and running code

A new code block will be added by pressing “Add Code” or by typing a command followed by (escape) and pressing “B”.

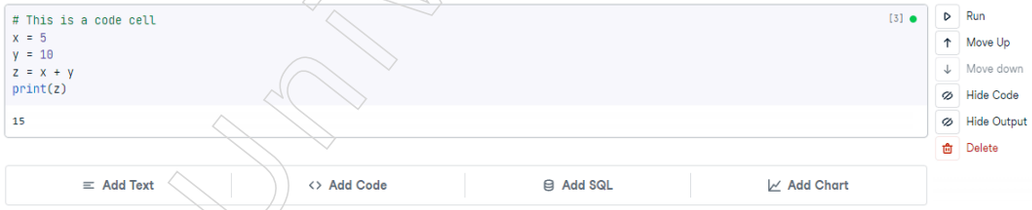


Write code in the cell just as you would in a script.



Jupyter Notebook console

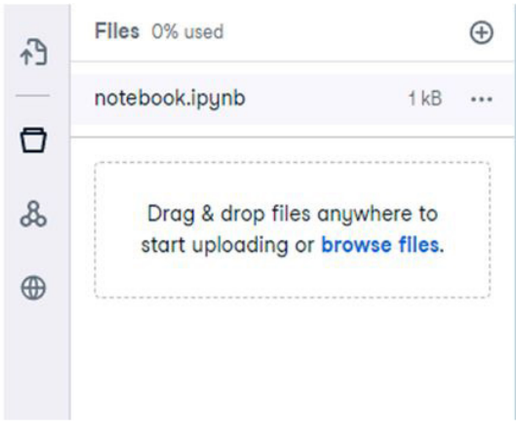
Pressing Run or CTRL/CMD+Enter runs the code and displays its output.



Jupyter Notebook console

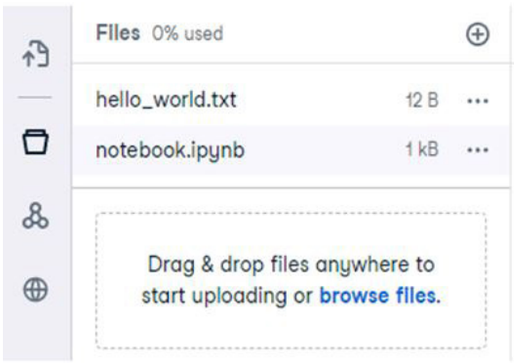
Reading and writing files

Pressing ‘Browse and upload files’ on the left-hand menu brings up the file system and pressing the ‘plus’ will allow you to upload a file from your local machine. Below, we have uploaded a simple text file called hello_world.txt.



Notes

Jupyter Notebook console



Jupyter notebook console

We can use the following code to open the file, add some text, then save a new file.

```
with open('hello_world.txt') as file:
    my_file = file.read()
```

```
print(my_file)
```

```
Hello World!
```

```
my_file = my_file + ' Goodbye World!'
print(my_file)
```

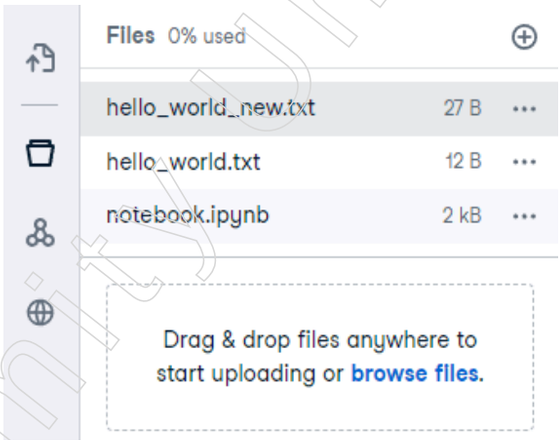
```
Hello World! Goodbye World!
```

```
with open('hello_world_new.txt', 'w') as file:
    file.write(my_file)
```

Add Text | Add Code | Add SQL | Add Chart

Jupyter Notebook console

You'll now see the new file in the file system and it will contain our updates.

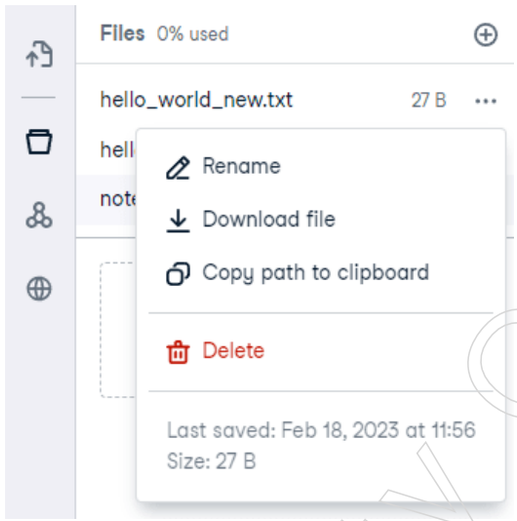


Jupyter Notebook console

Working with the File System

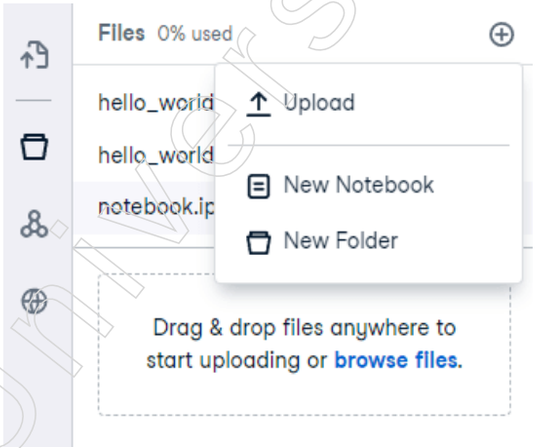
The steps to upload, update and create a new file are given above. Press the three dots in the file system and select download to download the new file.

Notes



Jupyter notebook console

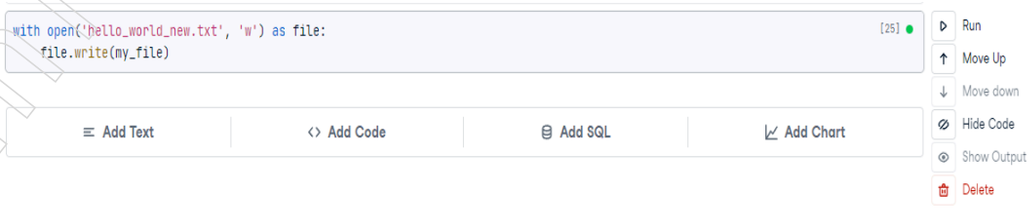
The plus button used to create new files can also be used to create fresh notebooks, which will have no cells or output.



Jupyter notebook console

Commanding cells

You may quickly reorder the cells using the move up and move down buttons, as shown in the image below.

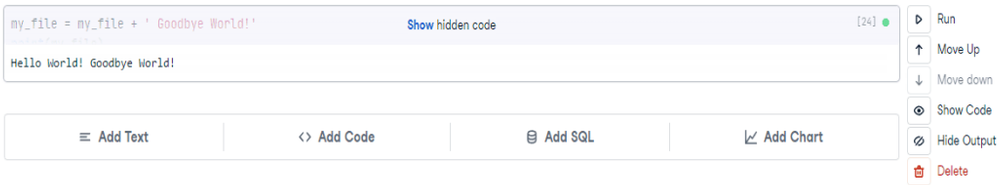


Jupyter Notebook console

It will rearrange your code. (Take note that running your code in the wrong order could cause it to malfunction!)

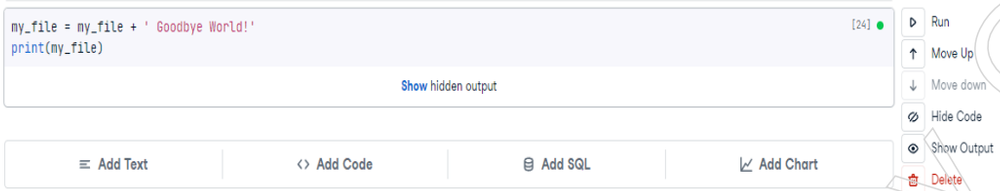
For long code blocks that you aren't currently working on, the Hide Code button will collapse and hide the code. It is also helpful if the readers of your analysis aren't interested in the specifics of the analysis and are simply interested in the outcomes.

Notes



Jupyter Notebook console

Similarly, the Hide Output button allows you to hide long outputs.



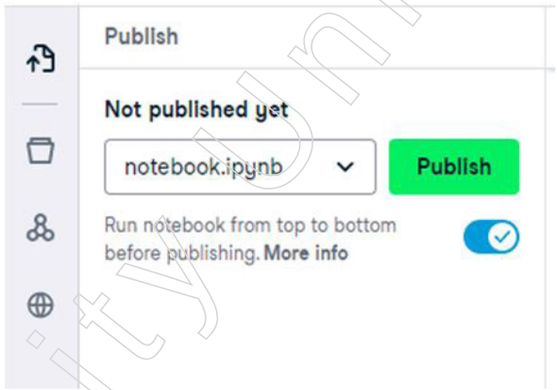
Jupyter notebook console

These buttons can also be used together to hide both code and output.

Publishing reports (Workspace only)

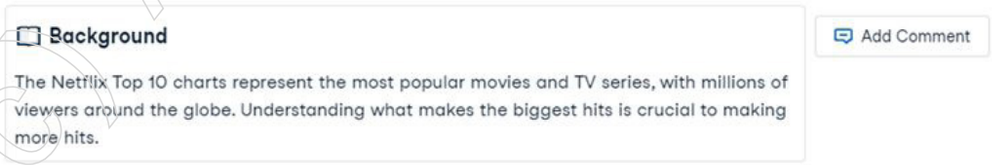
Your notebooks can be published as publications using Workspace. This is a fantastic opportunity to share your outstanding work and work with other data scientists.

Pressing the 'Publish' button on the side menu will allow you to share your notebook. Press publishes to share your notebook after that. Before publishing, it is a good idea to read the notebook cover to cover. Since most people read from top to bottom, this helps you examine your code and make sure it is readable.



Console

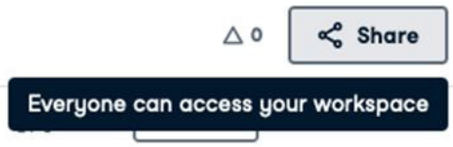
After you publish your notebook, other users can examine it and leave comments on specific cells. To other people, you can also do the same. This is a fantastic approach to start a conversation or comprehend a challenging piece of code. Here's a workplace example:



Console

Notes

Sharing Jupyter Notebooks (Workspace only)



Another helpful Workspace-only feature is the ability to share workspaces. You can share a public or private, access-controlled link that the recipient can execute independently because the notebook is hosted.

This is a wonderful approach to work together. Since data science is a broad and deep discipline, no one person can possibly know everything about it. For the greatest outcomes, whether they be effective code, attractive visualisations, or a precise model, data scientists must work together. Real-time collaboration using workspace enables numerous users to simultaneously edit a notebook.

Press the share button in the top right corner of your notebook to share it. Here, you can copy the link, set the notebook's privacy (if private) and public access settings and copy the link.

5.1.3 Jupyter Labs

A huge umbrella project called Project Jupyter, which aims to provide tools (and standards) for interactive computing using computational notebooks, includes JupyterLab, a highly extendable, feature-rich notebook creation programme and editing environment.

JupyterLab is a sibling of Jupyter Notebook and Jupyter Desktop, two additional notebook authoring programmes included in the Project Jupyter family. Compared to Jupyter Notebook, JupyterLab provides a more sophisticated, feature-rich and customizable experience.

The JupyterLab interface has a menu bar, a left sidebar that may be collapsed and a primary work area with tabs for documents and activities.

Menu bar

The top-level menus in the interface's menu bar list all the activities available in JupyterLab along with their keyboard shortcuts:

- File: Actions related to files and directories
- Edit: Actions related to editing documents and other activities
- View: Actions that alter the appearance of JupyterLab
- Run: Actions for running code in different activities such as notebooks and code consoles
- Kernel: Actions for managing kernels
- Tabs: A list of open documents and activities
- Settings: Common settings and an advanced settings editor
- Help: A list of JupyterLab and kernel help links

Left sidebar

The following features are accessible through clickable tabs on the left sidebar:

Notes

- File browser: A list of saved notebook documents and directories
- Data explorer: Browse, access and explore datasets and schemas
- Running kernels and terminals: A list of active kernel and terminal sessions with the ability to terminate
- Commands: A list of useful commands
- Cell inspector: A cell editor that provides access to tools and metadata useful for setting up a notebook for presentation purposes
- tabs: A list of open tabs

You can end your session in JupyterLab to stop the use of further resources. To end your session, first pick the power icon, then choose Shut Down from the popover that appears. After 12 hours without activity, notebook sessions end automatically.

Select the restart icon , which is immediately to the left of the power icon and then choose Restart from the popover that displays to restart JupyterLab.

Kernals

The language-specific processing tools used to process notebook cells are called notebook kernels. JupyterLab offers R, PySpark and Spark (Scala) language support in addition to Python. A notebook document's related kernel is launched when you open it.

The kernel executes a notebook cell and generates results, which may use a large amount of CPU and memory resources. Keep in mind that until the kernel is shut down, allocated memory is not released.

According to the table below, specific kernels are required for a specific set of features and functionalities:

Kernel	Library installation support	Platform integrations
Python	Yes	Sensei ML Framework
		Catalog Service
		Query Service
R	Yes	Sensei ML Framework
		Catalog Service
Scala	No	Sensei ML Framework
		Catalog Service

5.1.4 R Studio IDE

We can interact with R more easily with the help of RStudio, an integrated programming environment. Like the regular RGui, RStudio is thought to be more approachable. This IDE includes a lot of customization options, windows with many tabs and drop-down menus.

Three Windows will appear when initially launch RStudio. The fourth Window will by default be hidden. We may access this hidden window by selecting it from the File drop-down menu, New File and R Script, respectively.

Notes

RStudio Windows/Tabs	Location	Description
Console Window	Lower-left	The location where commands are entered and output is printed.
Source Tabs	Upper-left	Built-in test editor
Environment Tab	Upper-left	An interactive list of loaded R objects.
History Tab	Upper-left	List of keystrokes entered the console.
Files Tab	Lower-right	File explorer to navigate C drive folders.
Plots Tab	Lower-right	Output location for plots.
Packages Tab	Lower-right	List of installed packages.
Help Tab	Lower-right	Output location for help commands and help search Window.
Viewer Tab	Lower-right	Advanced tab for local web content.

Installation of RStudio

There are versions of RStudio Desktop for Linux and Windows. On both operating systems, installing the open-source RStudio Desktop is quite easy. RStudio's licenced edition includes a few more features than the free source version. Let's look at the extra features in the RStudio licence version before installing it.

Factor	Open-Source	Commercial License
Overview	1) Access RStudio locally	All of the features of open source are include with
	1) There is a commercial license for organisations which are not able to use AGPL software.	
	2) It provides access to priority support.	
	2) Code completion, syntax highlighting and smart indentation	
	3) Can execute R code directly from the source editor	
	4) Quickly jump to function definitions.	
	5) Easily manage multiple working directories using projects.	
	6) Integrated R help and documentation.	
	7) Provide interactive debugger to diagnose and fix errors quickly.	
	8) Extensive package deployment tools.	
Support	It supports for community forums only.	1) It supports priority email. 2) It supports for an 8-hour response during business hour.
License	AGPL v3	RStudio License Agreement
Pricing	Free	$995/year

5.1.5 Watson Studio

Data exploration, model creation and training are all accelerated and data science operations are scaled throughout the data science lifecycle with the aid of IBM Watson

Notes

Studio, a data science and machine-learning service. It is an end-to-end solution that makes it possible for developers and data scientists to work together effectively and to maintain the whole lifecycle of a model contained in a single environment.

Watson Studio provides the environment and tools required for you to collaborate with colleagues to address business challenges with data. To analyse and visualise data, clean and shape data, ingest streaming data, or construct and train machine learning models, you can select the tools you require.

Watson Studio's architecture is centred on the analytics project. Data scientists and business analysts organise resources and analyse data using analytics initiatives. This image shows how an analytics project for Watson Studio is organised and how interactions take place.

Watson Studio service:

- Data Refinery: Prepare and visualize data.
- Jupyter notebook editor: Code Jupyter notebooks.
- JupyterLab IDE: Code Jupyter notebooks and Python scripts with Git integration. Other project tools require additional services. See the lists of supplemental and related services.

IBM Watson Services

1. Watson Studio enables you to prepare and analyse material in a single integrated environment while training, deploying and managing your AI models.
2. Through dynamic data policies and standards, Watson Knowledge Catalogue promotes cooperation and turns data and AI into a trusted enterprise resource.
3. Watson Assistant enables you to create chatbots and virtual assistants for a variety of platforms, including robotics, messaging apps and mobile devices.
4. Using the most sophisticated cloud-native insight engine in the world, Watson Discovery uncovers hidden value in information to provide answers, track trends and correct problems.
5. Watson IoT Platform helps to make and maintain an efficient IoT infrastructure
6. Watson Speech to Text (STT) helps convert audio/speech to text.
7. Watson Text to Speech (TTS) helps convert text to audio/speech.
8. Watson Language Translator helps translate between different languages.
9. Watson language Classifier helps you classify the natural languages getting used.
10. Watson's language Understanding helps you understand natural languages.
11. Watson Visual Recognition allows you to tag, classify and train visual content using machine learning rapidly and precisely.
12. Watson Tone Analyser allows you to determine whether a person is furious, joyful, or whether the music is nice or not by analysing the tone of their voice.
13. You can learn more about personality traits with the aid of Watson Personality Insights.
14. Data Refinery offers you custom models that identify items and relationships specific to your industry, enabling you to teach Watson the language of your domain.
15. Watson Machine Learning gives you the ability to create, train and use machine learning and deep learning models using your own data.

Notes

16. Deep Learning helps you build deep learning models.
17. Watson Contract workflows are streamlined by Compare and Comply to reduce time spent, increase accuracy and simplify contract control.

Advantages Of Using IBM Watson

1. Watson gives you complete control over your data, models, learning and API, which are the cornerstones of your competitive edge.
2. Due to Watson's tremendous learning capacity, it can learn more from less.
3. Watson was once solely accessible through IBM Cloud, but it is now usable by any company powered by the cloud. As a result, clients are free to install AI wherever their data is located and are not forced to work with a single vendor going forward.
4. You'll find fresh trends and insights using Watson. Project possible future outcomes.

Disadvantages of using IBM Watson

1. IBM Watson is only available in English. Thus, it limits the areas of use.
2. It does not immediately process structured data.
3. Despite the volume of data increasing, there are still little resources available to meet the demands.
4. In the case of IBM Watson technology, maintenance is a significant concern.

Barriers To Adoption of IBM Watson

1. High Switching Cost.
2. It takes time and effort to integrate IBM Watson and its services into a company.
3. IBM Watson is targeted towards bigger organisations that can afford Watson.
4. It takes time and effort to teach Watson to use it to its full potential.

5.1.6 Other IBM Tools

The practise of data science is not without difficulties. There are several tools, practises and frameworks to pick from, fragmented data, a shortage of data science expertise and strict IT standards for implementation and training. Additionally, operationalizing ML models with ambiguous accuracy and elusive predictions presents challenges.

In other words, you gain the capability to implement data science models on any cloud while fostering confidence in AI results. Additionally, ModelOps will allow you to oversee and manage the AI lifecycle, while prescriptive analytics will help you make better business decisions and visual modelling tools will speed up time to value.

The first consideration in a data science project is where and how we will get our data. What's its origin? It will be structured or unstructured, right? etc. It can be gathered using a variety of methods, including web scraping, using sensors, configuring APIs and querying databases, to name a few.

It is crucial to note that depending on the analysis we want to conduct, it is essential to make sure that the data is current and that our environment has adequate room to keep it.

The following procedures are to clean, organise and store the raw data once we receive it. It is quite likely that we will get unexpected or erroneous results if we don't

Notes

sanitise our data. We can employ an ETL tool to carry out these processes. ETL (Extract-Transform-Load) is a procedure that lets us combine data from several sources, arrange it and store it in a single location.

ETL solutions are sold by a variety of companies; examples include IBM's InfoSphere DataStage ($), Amazon's AWS Data Pipeline ($) and Microsoft's Azure Data Factory ($). These technologies all include user-friendly visual user interfaces that help non-programmers determine the information flow. They can frequently be coupled with their own cloud platforms and other services.

However, these tools occasionally include functionality that go beyond what is required to address a problem. We can discover tools like Apache Camel, Apache Nifi, Apache Airflow and Logstash on the open-source side. Open-source software frequently have fewer features and can be a little trickier to use.

They occasionally provide us the option to modify their code so we can tailor it to the requirements of our application. However, if what is available in the market doesn't suit your requirements, there is always the option of developing your own ETL using your favourite programming language.

Loading the data is the final step of the ETL process, which raises the following inquiries: Where will we put it in storage? How will we keep it safe? These days, we can choose from a variety of databases and locations, including internal databases and cloud databases. The following table is a list of several well-known databases:

SQL		NoSQL	Data model	Licensing
Cassandra		X	Column-family	FOSS
MongoDB		X	Document	SSPL
Redis		X	Key-value	BSD
MySQL	X		Relational	GPL
PostgreSQL	X		Relational	PostgreSQL License
DB2	X		Relational	$$
Oracle DB	X		Relational	$$
SQL Server	X		Relational	$$

It's critical to remember that there are other methods for storing data besides databases. Large data sets can be stored and processed using platforms like Hadoop. What instruments can we use to conduct some analysis and processing after all the data has been put in the repository?

If programming is not your strong suit, you could find Rapid Miner ($), Data Robot ($), Trifacta ($), Excel ($), SAS ($) and IBM Watson Studio ($) useful. These tools have numerous functionality that cover most application cases.

You might not be able to finish the analysis with them, though, if your situation does not fit any of the accessible possibilities. However, if you are proficient in programming, Python and R (RStudio) are the two technologies that are most widely used for Data Science projects.

It is advised to utilise Python for online analysis and method implementations and R for offline analysis. When we are interested in applying machine learning, tools like IBM SPSS ($$), Matlab ($$), BigML ($$), Tensor flow and Weka can be useful. Let's look at Spark, BigQuery and Hadoop where processing speed is important.

Notes

Even if the previous procedures are correctly completed, a project's success cannot be guaranteed if the results are not presented in a sufficient and understandable fashion. Understanding the audience, their interests and their background is essential when presenting information to select the best method for visualising and communicating the outcomes. Tableau ($), ggplot2 (works with R) and Matplotlib (works with Python) are some common visualisation tools used in data science projects.

The optimal tool for you and your project will ultimately rely on a wide range of factors, but it is always helpful to have a starting point. If you are unfamiliar with R or Python, you should add them to your list of things to learn along with various methods for data analysis. Of course, this is just the beginning and there are a lot more necessary abilities to learn.

Tools will assist us solve the problem (save data, apply algorithms, etc.), but to get the desired results, the project participants must also have a solid grasp of data bases management, statistics, data modelling, etc.

5.1.7 Python

You can use the free and open-source Jupyter Notebook web tool to create and share documents that contain live code, equations, visuals and text. Employees of Project Jupyter are responsible for Jupyter Notebook maintenance.

Python is a well-liked programming language that may be used to create software and websites, automate processes and analyse data. Since Python is a general-purpose language, it can be used to develop a variety of applications and isn't intended to solve any specific issues. Due to its versatility and beginner-friendliness, it is currently one of the most popular programming languages.

According to the 2022 Developer Survey by Stack Overflow, almost half of respondents use Python in their development work, making it the fourth most popular programming language overall. The results of the survey also revealed that Python and Rust are the two most desired technologies, with 18% of developers who aren't already using Python stating that they are eager in learning it.

What can you do with python? Some things include:

- Data analysis and machine learning
- Web development
- Automation or scripting
- Software testing and prototyping
- Everyday tasks

Why is Python so popular?

Python is popular for several reasons. Here are some details on what makes it so flexible and user-friendly for programmers.

- Because of the clear grammar, which mimics real English, it is simple to read and comprehend. As a result, projects can be built and improved more quickly.
- It can be modified. Python is useful for many different things, including web development and machine learning.
- It's user-friendly for beginners, making it well-liked by beginning programmers.
- Since it is open source, anyone may use and distribute it without charge, even for commercial purposes.

Notes

- ❖ Python has a substantial and growing library of modules, or groups of code created by outside developers to improve Python's capabilities.
- ❖ Python has a sizable and vibrant developer community that adds to the language's library of modules and functions and serves as a valuable resource for other programmers. If programmers come into a problem, finding a solution is usually not too difficult because of the large support group; someone has almost certainly faced the same issue before.

5.1.8 Github

GitHub is a web-based platform for version control and collaboration for software engineers. Microsoft, GitHub's largest individual donor, bought the service for $7.5 billion in 2018. GitHub was founded in 2008 and employs the software as a service (SaaS) model of delivery.

It was built on the open-source Git code management system, which Linus Torvalds created to speed up software development.

Git is a tool for storing project source code and tracking all code alterations. It gives developers the means to manage possibly conflicting changes from different developers, allowing them to work on a project more successfully.

How does GitHub work?

Through the provision of a hosting service, a web interface for the Git code repository, collaboration management tools and other services, GitHub promotes social coding. Like a social networking website for programmers is the developer platform.

Members can follow one another, evaluate one another's work, obtain updates for certain open-source projects and engage in public or private communication.

Some key words that GitHub developers use include the ones listed below:

- Fork. A repository that has been cloned from one member's account to another member's account is referred to as a fork, also known as a branch. A developer can make changes using forks and branches without affecting the original code.
- Pull request. A developer can submit a pull request to the owner of the original repository if they want to share their updates.
- Merge. If, after reviewing the modifications, the original owner decides they want to add them to the repository, they can approve the changes and merge them with the original repository.
- Push. A programmer sends code from a local copy to the online repository in this case, which is the opposite of a pull.
- Commit. A solitary change to a file or collection of files is called a commit, or code revision. By default, commits are kept and interspersed into the main project; however, by using commit squashing, they can be integrated into a easier merging. Each time a commit is saved, a distinct ID is generated to allow team members to keep track of their individual contributions. A commit can be compared to a repository's snapshot
- Clone. A repository's local copy is known as a clone.

Benefits and features of GitHub

- ❖ GitHub makes it easier for developers to collaborate. Distributed version control

Notes

is another feature offered. on order to keep organised, development teams can collaborate on a single Git repository and track changes as they happen.

- In addition to the well-known SaaS offering, GitHub also provides an on-premises version. GitHub Enterprise supports numerous third-party apps and services in addition to integrated development environments and continuous integration tools. Compared to the SaaS version, it provides more security and auditability.
- Other products and features of note include the following:
- GitHub Gist allows users to exchange notes or other code.
- GitHub Flow is a simple, branch-based procedure for deployments that get updated often.
- GitHub Pages are static websites used to host projects that directly pull data from a person's or group's GitHub repository.
- GitHub Desktop allows users to browse GitHub from desktops running Windows or Mac OS X instead of visiting GitHub's website.
- GitHub Student Developer Pack is a free developer tool provided for students. It offers access to GitHub, programming tools and cloud resources.
- GitHub Campus Experts is a programme that students can utilise to establish technical communities and become leaders at their schools.
- GitHub CLI is an open-source, free command-line programme that allows users to access GitHub services like pull requests from their local terminal. By doing away with the necessity to switch contexts while coding, this capability streamlines processes.
- GitHub Codespaces enables users to access popular programming languages and tools through a cloud-based development environment. Users have a limited amount of free time in the coding environment before it transitions to a premium pricing structure. It operates in a container.

GitHub use cases

GitHub is used to store, track and collaborate on software projects in several different contexts:

1. Businesses employ GitHub as a version control system, enabling members of the development team to follow updates to the source code as developers work together on it. Project management is made simpler by enabling multiple programmers to work on a project at once and ensuring that everyone is using the most recent version of the code.

   Additionally, it enables developers to refer to earlier versions if necessary. Because code is kept in a single area, GitHub makes it possible for developers to share code. Because it is a standardised way to store code, GitHub Enterprise also aids with regulatory compliance.
2. Programming instructors and students utilise GitHub in a variety of ways. Teachers and students have access to a variety of inexpensive tools with the Student Developer Pack. The platform is used by students to hold virtual events, engage on creative development projects and study web development.
3. Open-source software developers GitHub can be used to share projects with anyone who want to utilise or work on the software. To find flaws in suggested

Notes

code before modifications are finalised, developers' network, communicate and pitch their work to other developers in real time.

GitHub is categorised as a social media platform because of its networking and collaboration features; in the repository notes, it frequently links to other community websites like Reddit. Additionally, users can download programmes from GitHub.

4. Nonprogrammers work on document-based and multimedia projects using GitHub as well. The platform's version control facilities are helpful for cooperation and it is easy to use. The Art of the Command Line, for instance, offers a thorough introduction to the command line.

   Electronic musician Aphex Twin created the experimental music production tool Sample Brain. And a selection of food dishes can be found in the Open-Source Cookbook.

Getting started on GitHub

To sign up for GitHub and create a repository, new users and beginners follow these steps:

1. Learn about the command line. Users communicate with GitHub using the command line. Working with GitHub requires knowledge on how to utilise it; tutorials and other tools are available to assist in this process. GitHub Desktop is a different option.
2. Install Git. The instructions on the Git website can be used to install Git for free. A command-line version of Git is also installed when GitHub Desktop is installed. On many Mac and Linux computers, Git is already installed.
3. Create an account. Create a GitHub account using an email address by going to the GitHub website.
4. Create a new repository. Upon arriving at the GitHub home page, click the plus sign, followed by sample repo. When prompted, identify the repository and give a succinct description. Include the project license, .gitignore template and README file. Then click Create repository after scrobbling to the bottom of the page.
5. Creating a repository is the first step to collaborating on code in GitHub.
6. The user's GitHub page should now display a bare repository. They can use the terminal's git init command to make a local clone of that repository.

5.1.9 SQL to Query Data

Structured Query Language is known as SQL. It is used to handle or alter databases in computer programming. To manage databases, SQL queries are used. In plain terms, a query is more akin to a question or a request.

Let's say you asked, "Could you please send me the Employee IDs of every employee in the Accounts department?" How many seats are reserved for the show, for example? Consequently, we are asking for information or questioning.

In general, in a SQL query, you ask databases to fetch (or retrieve) certain data. We utilise SQL, a popular language, to do database queries. When businesses have a tonne of data they wish to alter, they use it. You are allowed to use SQL if you keep your data in a relational database.

Structured Query Language is known as SQL. It is one of the main query languages

Notes

used for relational database management and data stream processing. We may access and modify databases using SQL. There are many uses for SQL in the modern world.

Why Use SQL Query?

To search for or retrieve data from databases, use SQL Query. The following actions are possible with the SQL query-

- Use a SQL query to build a new database and add data to it.
- To obtain (or fetch) data from the database, use a SQL query. Additionally, to update or edit the database's current data.
- Using a SQL query, remove or drop the data or table from the database. After that, we can also construct a new table
- Using the SQL query to modify the tables', views' and procedures' permissions. Additionally, to develop stored procedures, views and functions.

How To Write SQL Query?

The SELECT clause, which permits choosing the data to be shown, serves as the foundation of a query in SQL Server.

An SQL SELECT statement retrieves records from a database table according to clauses (for example, FROM and WHERE) that specify the criteria based on which our data will be selected. The syntax for the SQL SELECT statement is:

column1, column2

FROM table1, table2

WHERE column1 = 'xyz' and column2 = 'abc';

SELECT

In the above SQL statement:

- The one or more columns to be obtained from the database are specified in the SELECT clause. A comma and a space should be used between the names of the columns when specifying multiple columns. However, we can use the wildcard * (an asterisk) to retrieve all columns.

SELECT * FROM ....

- The FROM clause designates which table(s) should be queried. When providing multiple tables, we can separate the table names with a comma and a space, for instance, FROM Names, Addresses, Phone_Numbers, where Names, Addresses, etc. are the table names.
- The WHERE clause only chooses rows that have the provided value in the specified column. WHERE enables you to narrow down a search query. Typically, the value is encased in single quotes (for instance, WHERE colour = 'teal').
- The statement terminator is the semicolon (;). However, you can omit the semicolon (;) if your SQL query statement is only one line long. However, if your query has many lines, it must be mandatory. In general, it is preferable to place a semicolon after each statement that ends a SQL query.

5.2 Major Visualisation Packages

The development of data visualisation technology has made it simpler for data visualisation designers to create visual representations of large data sets. When a

Notes

designer is working with data sets that comprise hundreds of thousands or millions of data points, automating the process of creating a visualisation, at least in part, greatly simplifies their job.

5.2.1 MS Power BI

Power BI, a business intelligence tool developed by Microsoft, processes and visualises raw data to generate insights that may be applied to decision-making. It brings together business analytics, data visualisation and best practises to support an organisation's decision-making using data. Because of the capabilities of the Power BI platform, Gartner recognised Microsoft as the Leader in its "2019 Gartner Magic Quadrant for Analytics and Business Intelligence Platform" in February 2019.

When a designer is working with data sets that comprise hundreds of thousands or millions of data points, automating the process of creating a visualisation, at least in part, greatly simplifies their job.

Why Power BI?

Following are the reasons why Power BI is so popular and needed in the BI domain:

1. Access to Volumes of Data from Multiple Sources
2. Interactive UI/UX Features
3. Exceptional Excel Integration
4. Accelerate Big Data Preparation with Azure
5. Turn Insights into Action
6. Real-time Stream Analytics

Advantages Of Power BI

1. User-friendly interface: Users of Power BI can readily see and analyse data thanks to the program's user-friendly interface.
2. Data integration: Users may quickly combine data from a variety of sources, such as Excel, SQL Server and cloud-based sources like Azure and Salesforce, using Power BI.
3. Customizable dashboards: Users can design dashboards and reports that are uniquely theirs to display data in a way that makes sense to them.
4. Real-time data: Power Real-time data processing is supported by BI, allowing users to view current data in their dashboards and reports.
5. Collaboration: Power BI makes it simple to collaborate on data analysis projects by allowing users to share their dashboards and reports with others.

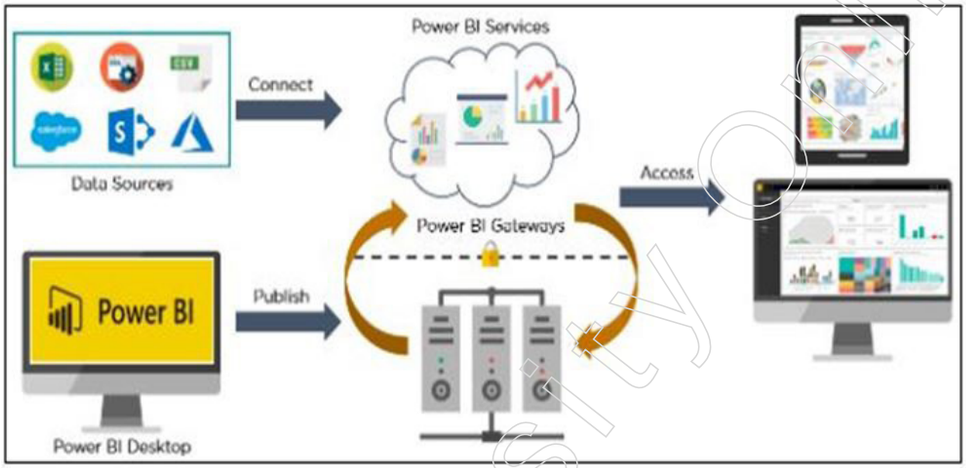
Disadvantages Of Power BI

1. Limited data processing capabilities: Power BI may have trouble processing huge datasets or sophisticated queries because it is not intended for intensive data processing.
2. Limited customization options: Power BI provides a variety of customization choices, however users could find that their capacity to produce really original visualisations and reports is constrained.
3. Cost: Users who want more features or storage space from Power BI may have to pay for it since it is not a free service.

Notes

Power BI Architecture

Azure serves as the foundation for the Power BI service. There are numerous data sources that Power BI may connect to. Power BI Desktop allows you to create reports and data visualisations based on the dataset. For continuous data for reporting and analytics, on-premises data sources are linked to the Power BI gateway.



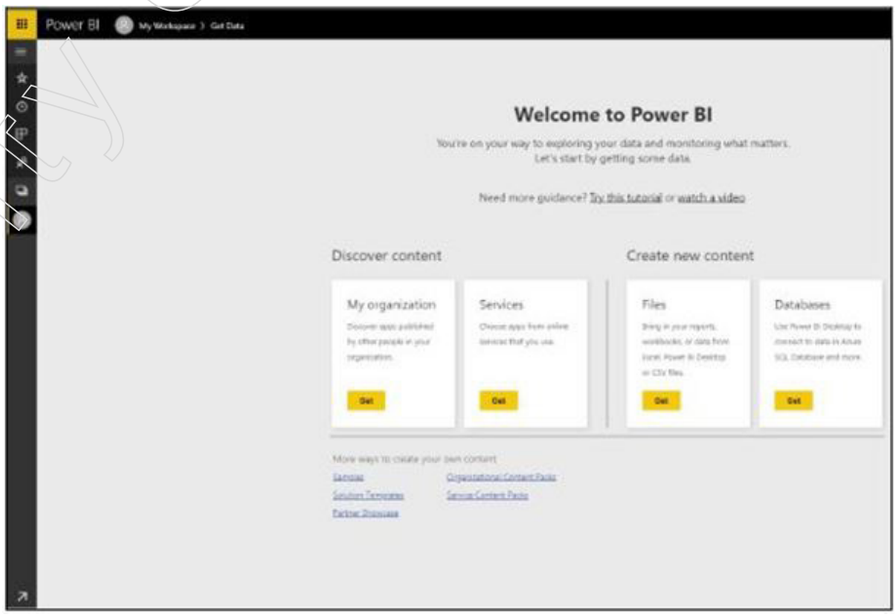
Power BI Architecture

The cloud services that are utilised to produce Power BI reports and data visualisations are referred to as Power BI services. You can stay connected to their data from anywhere with Power BI mobile apps. There are Power BI apps for Windows, iOS and Android.

Power BI Service

The Software as a Service (SaaS) component of Power BI is called Power BI service. It also goes by the name Power BI Online. You must sign into the Power BI service in order to access the service.

Here is how the home page of Power BI Service looks like once you log in:



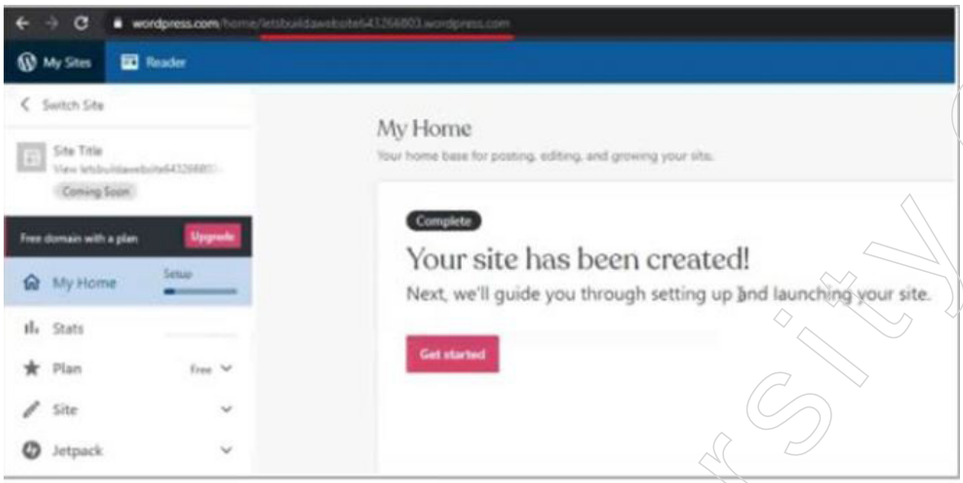
Power BI console

Notes

It enables you to access data, produce reports and dashboards and query your data.

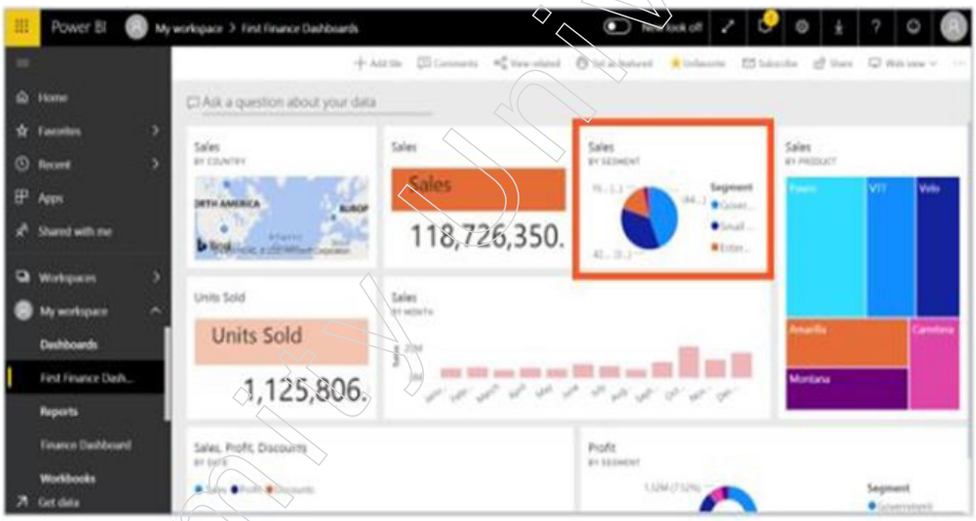
Power BI Dashboard

The Power BI Dashboard allows a story to be told on a single page. Reports are used to create the visuals on a dashboard and each report is built using a distinct dataset. A one-page dashboard is known as a canvas. The Finance Dashboard shown below was posted to the Power BI Service.



Power BI console

The dashboard's visual representations are known as Tiles and report creators have placed them there.



Power BI console

You may create a number of reports on Power BI Desktop. These reports can be published on the Power BI dashboard thanks to the Power BI service. A Power BI report created on Power BI Desktop can be uploaded to Power BI Service by choosing the "Publish" button.

Components of Power BI

1. Power Query
2. Power Pivot

Notes

3. Power View
4. Power Map
5. Power BI Desktop
6. Power Q&A

Features of Power BI

Following are some of the features of Power BI -

1. Power BI Desktop
2. Stream Analytics
3. Multiple Data Sources
4. Custom Visualisation

Who uses Power BI?

When a designer is working with data sets that comprise hundreds of thousands or millions of data points, automating the process of creating a visualisation, at least in part, greatly simplifies their job. Users of Power BI aren't simply limited to data experts like data scientists or data engineers; they can include a wide variety of business users. The platform is designed to make it simple for non-technical people to create reports, alter data and do in-depth data analysis activities.

Nonetheless, some of the most common analyst positions that use the platform daily include the following:

1. Business analysts
2. Business intelligence analysts
3. Supply chain analysts
4. Data analyst

5.2.2 Tableau

The robust and rapidly expanding data visualisation tool is Tableau. Tableau is a business intelligence platform that enables us to evaluate raw data visually in the form of graphs, reports, etc.

Example: - You may use Tableau to examine any data you have, including Big Data, Hadoop, SQL, or cloud data and to visualise the data in a visual format.

With Tableau, data analysis is completed extremely quickly and worksheets and dashboards are used to build the visualisations. Any expert can comprehend the data produced by Tableau.

The Tableau software is completely non-technical and non-programming. Tableau makes it simple and quick to create visual dashboards.

Why use Tableau?

Here are some reasons to use Tableau:

1. Ultimate skill for Data Science
2. User-Friendly
3. Apply to any Business

Notes

4. Fast and Easy
5. You don't need to do any Coding
6. Community is Huge
7. Hold the power of data
8. It makes it easier to understand and explain the Data Reports

Features of Tableau

1. Data Blending: Data mixing is Tableau's main feature. It is used when combining pertinent facts from multiple data sources that you want to analyse collectively in a single view and represent as a graph.
2. Real-time analysis: When the Velocity is high and real-time data analysis is challenging, real-time analysis enables users to easily comprehend and evaluate dynamic data. Tableau's interactive analytics can assist in recovering important information from rapidly moving data.
3. The Collaboration of data: Data analysis is not a lonely endeavour. Tableau is designed for collaboration because of this. Members of the team can distribute data, conduct follow-up research and send simple visualisations to those who could benefit from the information. Success depends on ensuring that everyone can understand the information and make informed decisions.

Usage of Tableau software are listed below:

- ❖ Business Intelligence
- ❖ Data Visualisation
- ❖ Data Blending
- ❖ Data Collaboration
- ❖ Query translation into visualisation
- ❖ To create no-code data queries
- ❖ Real-time data analysis
- ❖ To manage large size metadata
- ❖ To import large size of data

This data visualisation tool has been utilised by the business intelligence sector ever since it was first developed. Tableau is widely used by businesses like Amazon, Walmart, Accenture, Lenovo and others.

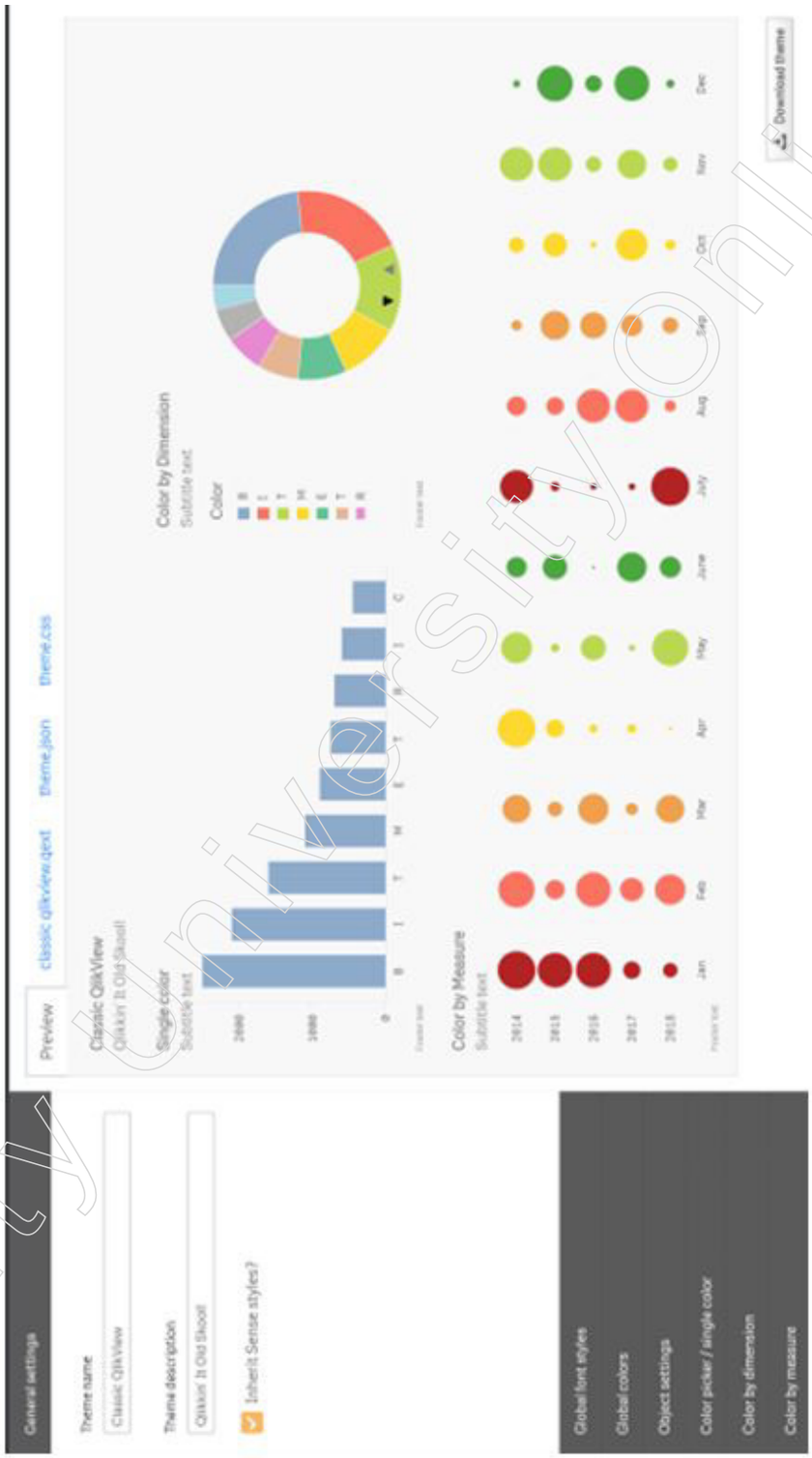
5.2.3 Qlik Sense

A full data discovery and analytics platform is Qlik Sense. Users may model and manage data, produce visualisations, layouts and stories using its cutting-edge interface. But what makes Qlik unique?

Its associative engine stands out. The analysis techniques used by the other vendors in this series are all query-based, which, according to Qlik, limits the results to linear and specified study of certain subsets of data. However, customers can view the data from any angle thanks to its engine.

Everything sounds fantastic, but Gartner warns potential customers to exercise caution given the low market momentum compared to its rivals.

Notes



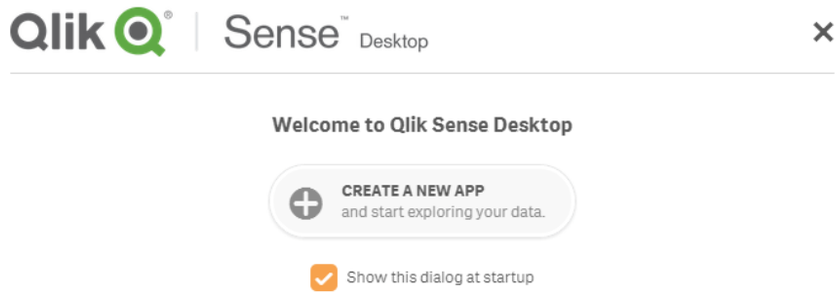
Qlik sense console

The self-service analytics tool is Qlik Sense. It enables individual users to build customised, interactive data visualisations, reports and dashboards easily and quickly from a variety of data sources. Even though Qlik Sense has the same robust engine as Qlik View, its appearance and feel are distinct.

Notes

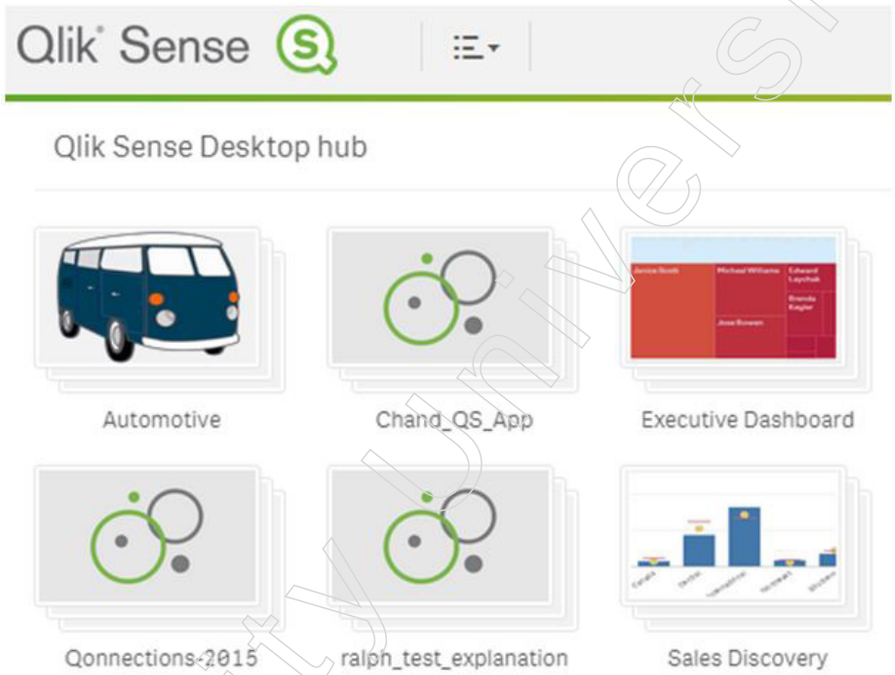
Qlik Sense features:

1. Free for internal company use and personal use. You may get Qlik Sense and begin using it by downloading it.
2. The application Qlik Sense desktop is incredibly dynamic and simple to use.



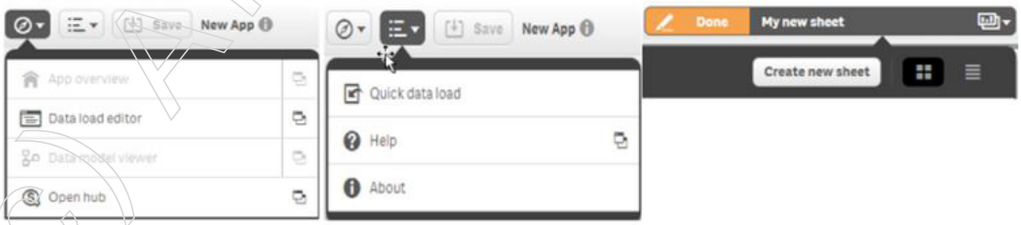
Qlik sense console

3. Sense hub is used to organise applications. Each app can contain multiple sheets



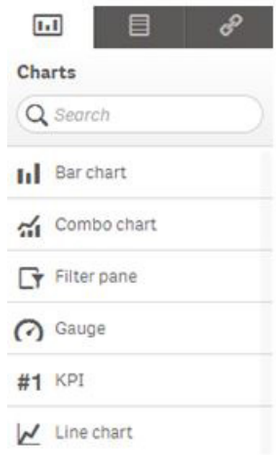
Qlik sense console

4. Data load can be as easy as simple “drag and drop” of files
5. Easy navigation

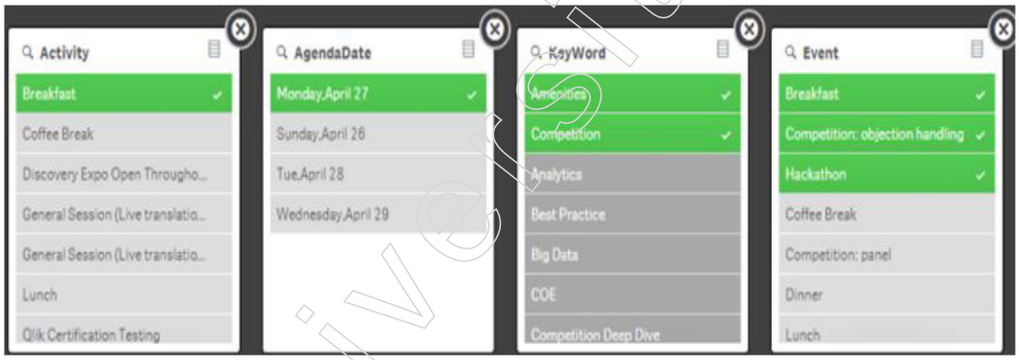


6. Easy to build charts by simple drag and drop

Notes



7. Functionality to create story
8. Global selectors available by default



9. Ability to combine your Qlik View scripts to Qlik Sense application
10. Powerful API's provide added extensions to the functionality

5.2.4 Klipfolio

Klipfolio is a web programme that is cloud-based and aids in your success with data. By comprehending, visualising and monitoring the KPIs and metrics that matter most to your business, you can succeed.

If you utilise Klips, PowerMetrics, or both, Klipfolio can assist you in making prompt, wise choices that will benefit both you and your company. Let's describes the steps in data journey:

1. Discovering

   With Klipfolio, you can access data from cloud applications, files on your computer, cloud file sharing services, SQL databases and email attachments.

   Not sure how to construct a data query? Not to worry! The most often requested data points are included in the pre-built data sources that Klipfolio offers. If you use PowerMetrics, you will have access to instant metrics, where you can connect to your data and produce a metric visualisation simultaneously in just a few simple steps.

   Klips users can benefit from our pre-built data sources. over instance, you can use our Facebook "Daily Reach" pre-built data source to receive a list of all the people who have seen material from your company's Facebook page overall, by day (over the past 90 days).

Notes

Want to make the custom-building process more efficient and straightforward? Edit your data sources using the Modeller before developing personalised metrics and Klips. It is simpler to comprehend and work with your data if you model the sources of your data.

2. Visualizing

Display your data! The time has come to visually exhibit your data, for instance, as a table or a bar chart. In Klipfolio, you may visualise your data in a variety of ways, including as metrics, individual Klips, or a dashboard of all of them.

- Metrics: A metric is a value that is tracked over time, such as profit or recurring revenue monthly. You can continuously access both current and historical information when you track the metrics that are most important to your company. This gives you the information you need to make the best decisions possible at the right moment. Use our quick and simple instant metrics or build your own unique measures.
- Metric Dashboards: Your linked metrics are displayed in metric dashboards. All our dashboards are completely customisable, enabling you to design one from scratch or add a pre-built dashboard template in order to create one that is unique and catered to your needs and audience. With this flexibility, you may successfully shape your data to enlighten others and show progress towards your business goals.
- Klips: A Klip is a graphical display of data in the form of a pie chart, bar chart, or gauge. Klips are dynamically updated dependent on the refresh rate of the connected data source and are pre-populated with your data. You can create custom Klips using the Klip Editor or use pre-built Klips from the Klip Gallery.
- Klip Dashboards: Your Klips are shown on a dashboard, which you can share with people or groups or put on big screens throughout the office. Dashboards are an excellent method to gather related content in one location. Check out our pre-built dashboards in the Dashboard Gallery or create your own dashboard from scratch.

3. Sharing

Make sure everyone is discussing your data! Share your analytics, Klips and dashboards with others to foster collaboration. Sharing data trends and outcomes keeps everyone on the same page and moving towards shared objectives. Additionally, it promotes conversation and significant action.

Klipfolio offers a variety of sharing options, including the ability to share URLs to dynamically updated views of your data as well as PDFs of your favourite dashboards that you have downloaded and shared. In your Klipfolio account, you may share with individuals, groups, or the entire staff at once by projecting your dashboards, metrics and Klips on sizable screens.

4. Taking action

What do you do with your data now that you have it? Most essential, constantly monitor your data! Evidence demonstrates that organisations who track their data consistently and everyday see trends and movements that aid them in making quick decisions.

You can create performance indicators in Klips that show changes in your data or perform historical comparisons of your important metrics across various time periods.

Notes

5 key benefits

We really value and utilise Klipfolio's five unique features in our service business.

1. Custom Styling
2. Roles, Groups and Rights
3. Data connections
4. A powerful formula editor
5. The API and SSO

5.2.5 Looker

Looker employs the Data Modelling Language (DML) and has a built-in framework. You may analyse data in a highly effective and beneficial way with Looker. Using Looker, you can quickly connect to many data sources and create customised dashboards, KPI dashboards, etc.

Looker offers tools to support a variety of data experiences, including contemporary BI and embedded analytics as well as integrating workflows and creating unique apps. Looker enables you to access the most recent version of your company's data, no matter where it is stored. It is a one-stop shop for business intelligence, analytics, visualisation and data management.

Features of Looker

1. Break down barriers to insight

   By quickly filtering to precise slices of data from the dashboard, you can obtain the crucial insights you require. Additionally, Slack allows you to start data in every chat and gives you instant access to the solutions you need.

2. Increase performance and optimize costs

   You can enhance performance, save costs and more effectively manage enterprise-scale deployments using Looker's robust features.

You can open up new categories of data experiences and quicken development processes by using prebuilt UI components. Looker increases your product's competitive advantage and accelerates revenue growth for a reasonable price.

Apart from this, we have many other new features. A few of them are listed below:

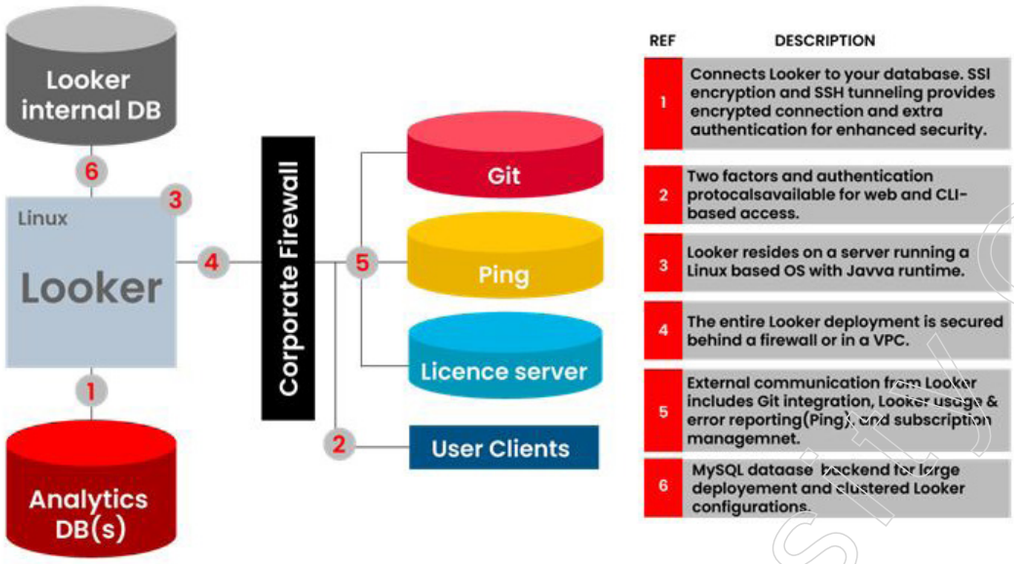
- Integrated end-to-end multi-cloud platform: You can carry out data analysis and visualisation using on-premises databases, Google Cloud, AWS and Azure.
- Common data: With effective data modelling that abstracts underlying data at any scale and produces a uniform data model for the entire organisation, business intelligence may be operated by everyone.
- Embedded data experiences: Looker offers embedded analytics that are quick to benefit from and customizable.
- Augmented analytics: Add cutting-edge machine learning, artificial intelligence and sophisticated analytical capabilities to Looker's business intelligence.
- Tailored data applications: Create data-centric applications for a variety of industries, including supply chain logistics and sales assistance, using embedded machine learning and interactive data visualisations.

We recommend you watch this Looker Tutorial Video learn Looker easily.

Notes

Looker Architecture

The Looker architecture consists of a Linux server with the connections listed below:



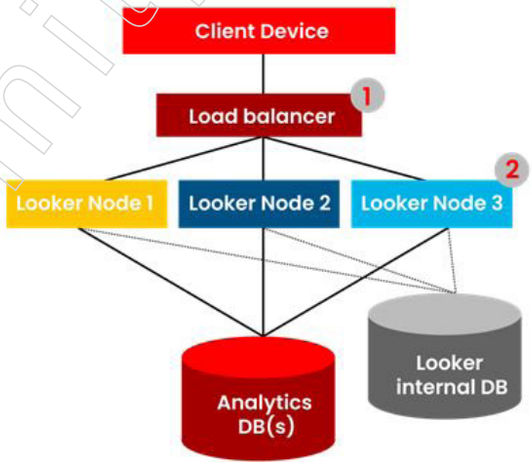
Looker Architecture

In an AWS VPC, Looker can be customer hosted. The amount of work required to install, configure and administer Looker apps and other IT tasks you undertake connected to Looker apps is decreased when using a Looker-hosted instance. Keep in mind that the hosting of Looker applications is unrelated to the location of data. Data is never copied to the Looker instance; it is always stored in the database.

Additionally, Looker requires outbound network access to authorize, backup Git, email relay and license checks.

For the internal databases of the Looker programme, HyperSQL is selected as the in-memory database by default. Performance problems result from the database's size growing in busy environments. Users must switch to a MySQL database backend instead of the HyperSQL database to get around these deployments.

Additionally, users can deploy a cluster of Looker nodes behind a load balancer, as indicated in the image below, to achieve effective traffic management on Looker:



Nodes

Notes

In a clustered system behind a load balancer/proxy server, various looker nodes can be deployed. Each node uses HTTPS to interact with the other, utilising additional security protections.

The suggested course of action for a production instance with a 100% uptime would be this.

What is a LookML?

In a SQL database, dimensions, calculations, aggregates and data relationships are described using the LookML language. Looker uses a model in LookML to construct SQL queries against a particular database.

5.2.6 Zoho Analytics

Zoho Analytics, a piece of business intelligence (BI) software from the Zoho Corporation, offers a user-friendly dashboard with several visualisation options. Through integration, it retrieves data from further Zoho applications.

Businesses can use Zoho Analytics to visually evaluate their data, produce unmatched data visualisations and unearth untapped insights. There is no longer any need for IT support or data analysts thanks to the data analytics integration, which provides measurements and analytical reports for the business data.

Together with Zoho Analytics partners, you can analyse large datasets, perform several analytical operations, import data, integrate that data, create custom reports and visually represent the results to uncover insights. By integrating with Zoho, which provides adaptable on-premises or cloud deployment options, businesses may base decisions on data.

Zoho Analytics Features

- Including business data from more than 250 sources. data from files, feeds, URLs and apps is synchronised.
- Seamless integration with other Zoho Suite applications as well as third-party applications via data connectors (Shopify, Google Analytics, Google Drive, Google Ads).
- Gathering and organising data. utilises ZohoDataPrep to shape large amounts of raw data.
- Deconstructing data silos to show the wider picture. combines data from several sources.
- Using analytics to facilitate collaboration. Share files with the organisation and change permissions.
- Telling stories using data. Create compelling narratives about the state of your business and support them with data visualisations.
- Making dynamic maps use of your data insights to visualise.
- Augmented analytics and artificial intelligence assistant: Zia, Zoho's AI helper, can deliver immediate information in the form of various KPIs with configurable visualisation options.
- Embedded analytics: Clients can purchase white label business analytics solutions from business consultants or developers.

Notes

- Interesting and insightful reports: You can simply customise your reports using a drag-and-drop interface.
- Simple and adaptable deployment: Zoho setup is quick and easy. Without additional costs, you can select an on-demand alternative or a cloud-based solution.
- Building portals, using them on various websites and using the BI app on mobile devices.
- Data security. Security is a priority for Zoho.

Zoho Analytics Pros

Zoho Analytics is one of Zoho's best applications. We rate it a solid A and highly recommend it.

- Easily integrates with more Zoho programmes, including Zoho Invoice, Zoho People, Zoho Books and Zoho CRM.
- Gives an immediate snapshot into many KPIs.
- Enables you to analyse your data in real time.
- Allows you to create and share visual dashboards.
- Combines multiple data sources into comprehensive snapshots of your business.
- Features beautiful reports.

Zoho Analytics Cons

Really, Zoho Analytics has no shortcomings. Maybe it's just a learning curve. But once you do, you're in business analytics paradise.

5.2.7 Domo

Domo, whose name is derived from the Japanese word for "Thank You," is a cloud-based solution created to combine data from various back-end systems to provide comprehensive views of organisational performance with filterable and customizable access to a range of key performance indicators (KPIs) and metrics.

The ability to examine real-time data on a variety of customisable dashboards and swiftly evaluate the data to enhance business performance and seize new possibilities is greatly facilitated by this. Domo provides a wide variety of data presentations that may be customised with complex graphs and multi-part widgets.

Data, technology and people are all combined on the Domo platform to create a digitally connected business. Discover how Domo's seven systems interact to provide the operating system for businesses.

5.3 Other Visualisation Packages

5.3.1 Infogram

Infogram is a programme for data visualisation that was created for marketers, media businesses and business strategists. With the help of this application, businesses may create extremely interactive visualisations. All team members have access to a collection of tools for creating infographics, graphs and charts. More than 500 maps, 35 charts, 20 pre-designed themes, numerous icons, a drag and drop editor and data export/import functionality are included in the tools.

Notes

On any device, users may create a completely responsive, polished presentation. Additionally, coworkers can simultaneously create and edit presentations. Users may easily customise presentations with the drag-and-drop editor and if more data is required, it can be imported into Infogram. The system may import data from both online and offline sources, whether it is kept on a computer or online. Infographics can be downloaded in HD resolution and in a variety of file formats, including PNG, JPG, PDF and GIF.

Anyone can create and share beautiful charts, maps, infographics, reports and presentations with the aid of infogram.

A tool for data visualisation called Infogram enables you to make interactive maps, infographics and charts. All throughout the world, newsrooms, marketing teams and students use it. With offices in San Francisco and Riga (Latvia), Infogram was founded in 2012 and quickly rose to prominence as the web's preferred data visualisation tool due to its simplicity, functionality and striking aesthetic. approximately 50 million people per month view the approximately 5 million infographics and visualisations that our users have generated.

5.3.2 Chartblocks

A reporting tool with customizable chart kinds is called Chartblock. On its SaaS platform, users can upload structured data and use the ready-made templates for mobile visualisation. Charts can be added to websites using the embedding code.

A tool for creating charts online is called chartblocks. It is the easiest chart builder app in the world, allowing you to create and share a chart in a matter of minutes.

No coding is necessary. Create a chart quickly with the user-friendly chart creator by selecting from a wide variety of chart types and then modifying it to suit your needs.

Draw information from virtually any source and even make charts that combine information from various sources. You will be guided step-by-step through the procedure by the data import wizard.

Use built-in social media sharing features to send your chart immediately to Facebook and Twitter by embedding it on any web page. With order to use your charts with Illustrator and other graphic design programmes, you can also export them as editable vector graphics.

5.3.3 Data Wrapper

Software for compiling and visualising data is called Datawrapper. It was created with the intention of including maps and charts in already published news articles that are posted on websites. Users can generate charts with only one click once the data has been loaded, which frequently requires copying and pasting right into the tool.

Simple interactive charts can be made using the user-friendly, open-source online tool known as Datawrapper. Datawrapper can be used to import your CSV dataset and produce a map or chart that can be embedded on a website. Maps, pie charts, line charts, vertical and horizontal bar charts and more are all possible to make. You may make three different visualisations with it:

- maps: choropleth maps, symbol maps and locator maps
- charts: from simple bar charts, line charts, column charts to arrow charts, scatterplots, population pyramids, etc.

Notes

- tables: with mini line charts, bar charts, images, Markdown, etc.

Datawrapper console

In 2012, when journalist Mirko Lorenz started working with developer Gregor Aisch, the concept for Datawrapper first came to him. Mirko was engaged in intensive data training at the time, but there was no easy way to make graphs, maps, or charts.

Together with Aisch, Mirko created a novel charting tool that would allow users to visualise their data rapidly and simply as charts or maps. To ensure that people anywhere can access this cutting-edge software, Datawrappers products are currently available in six languages: English, German, Spanish, French, Chinese and Italian.

For software developers, journalists and any design professional wishing to combine data into a certain format, Datawrapper is a potent tool. You can copy data from many sources, such Google Sheets, Excel, or the internet and utilise it in Datawrapper.

The user can then choose from a variety of maps or charts to represent this data. Additionally, it offers choices that let the appearance be altered to fit the style of other newsrooms or other sources. You can view the graphical representations created with Datawrapper on any BI tool and on devices with different screen widths.

Datawrapper offers a free plan in addition to its paid monthly options, which is an excellent choice for anyone who want to embed graphics on smaller or less popular websites. Small enterprises and students looking for a user-friendly data visualisation tool can consider Datawrapper. The top rivals of Datawrapper are Infogram, Plotly, Tableau and Charts.js.

Pros and Cons of Using Datawrapper

The following are a few of the benefits, as well as drawbacks, of using Datawrapper for your data visualisation needs:

Pros

- Datawrapper can work on any OS computer.
- It can be used online or installed onto a server.
- The basic version of Datawrapper is free.

Notes

- ❖ It can be used to create 19 types of charts as well as three kinds of maps.
- ❖ Because it provides web-based visualisation capabilities, Datawrapper users won't lose data if their computer crashes and can access data and charts from any location. All changes are automatically saved.
- ❖ It's easy to present data using Datawrapper's clear, beginner-friendly user interface. This software doesn't require that users have prior knowledge of coding or web design to begin creating data visualisations.
- ❖ Datawrapper offers a wide range of design defaults that can be used to streamline the visualisation process for those who are new to making data visualisations. These make sure that the data visualisation is understandable and appealing to the eye.
- ❖ Options for customising how data is presented are available. Users can upload images, alter the margins, choose a different colour scheme and modify the fonts.
- ❖ The information you upload to Datawrapper is private and belongs to your account only. No reader data is collected.
- ❖ Before exporting the data visualisation, customers can examine a preview to make sure it matches their expectations.
- ❖ Interactive maps may be produced using Datawrapper, which results in a more involved and active user experience.
- ❖ Datawrapper is open source.

Cons

- ❖ Datawrapper has limited flexibility when working with visuals.
- ❖ Some Datawrapper users have a difficult time customizing fonts and colours. In addition, adjusting colour in Datawrapper can be challenging for users.
- ❖ Any data that's uploaded when using the free plan will be transferred to the Datawrapper server's storage.
- ❖ Because it is open-source, security concerns can arise.

Five steps are typically included in the construction of an interactive chart. Find the dataset first (you may find several sample datasets in the first phase of the chart development process if you want to play with it). Next, purge the data (e.g., eliminate pointless rows, create a CSV file from it).

Upload it to Datawrapper third. The fourth step is to visualise the data by choosing the appropriate types of charts (such as a pie, bar, or line chart). Finally, obtain the embed code and post the chart on your website.

5.3.4 Google Charts

The goal of the charting library Google Charts, which is entirely composed of JavaScript, is to enhance online applications by providing interactive charting. It accommodates a variety of charts. In common browsers like Chrome, Firefox, Safari and Internet Explorer (IE), charts are drawn using SVG. VML is used to draw the visuals in IE 6's legacy version.

Features

Following are the salient features of Google Charts library.

Notes

- Compatibility − Effortlessly operates on all popular browsers and mobile devices including Android and iOS.
- Multitouch Support − supports multitouch on systems with touch screens, such as iOS and Android. Ideal for Android-based smart phones and tablets and iPhone/ iPad devices.
- Free to Use − Free to use for non-commercial purposes and open source.
- Lightweight − The basic library of loader.js is quite small and light.
- Simple Configurations − uses Json to create different chart configurations and it's incredibly simple to understand and use.
- Dynamic − Allows to modify chart even after chart generation.
- Multiple axes − Not restricted to x, y axis. Supports multiple axis on the charts.
- Configurable tooltips − Tooltip comes when a user hover over any point on a charts. googlecharts provides tooltip inbuilt formatter or callback formatter to control the tooltip programmatically.
- DateTime support − Specifically handle date and time. offers a variety of built-in options for date-based categories.
- Print − Print chart using web page.
- External data − Supports loading data dynamically from server. Provides control over data using callback functions.
- Text Rotation − Supports rotation of labels in any direction.

5.3.5 Fusion Charts

FusionCharts.NET is a charting framework for ASP.NET MVC, ASP.NET WebForms,. NET Core and.NET Standard that renders interactive charts using the FusionCharts JavaScript (HTML5) library.

FusionCharts.NET offers an object-oriented paradigm in which you may render charts using C# or VB, preventing you from having to write messy JavaScript and JSON code within your ASP.NET code. There are two modules included:

- Data Engine - FusionCharts.NET's data engine enables you to store data and work with it. As a result, before sending the data to the chart renderer, you can alter and optimise it. Only data in the form of a DataModel is accepted by the chart renderer in FusionCharts.NET. The following components make up the data engine:
- The data source class builds an object with the raw data and passes it to the DataModel.
- The transformed data is received and stored by the DataModel. You can give the chart renderer this parent DataModel in order to display the data.
- Data operations are also included in a data model and can be used to build numerous optimised data models. You can give the chart renderer any of these DataModels.
- Visualisation - After obtaining the data from the data engine, the visualisation module of FusionCharts.NET enables you to generate the chart with functional and aesthetic customizations.

5.3.6 Sisense

The Sisense data and analytics platform is made to make it simple to combine data

Notes

from all areas of the data landscape and turn it into embeddable, actionable analytics apps. Globally, businesses utilise Sisense to promote digital transformation and speed up innovation. Big names like GE, Nasdaq and Philips are among the innovative startups that the vendor represents. Regardless of whether data is stored on-premises, in the cloud, or a combination of both, Sisense gives users everything they need to successfully implement analytics application strategy today:

- Securely connect to cloud data warehouses and ingest data there to consolidate all your data into one location.
- Combine data from a variety of sources to get a full picture of your company.
- Use a combination of live and cached data models to manage resource usage and improve performance.
- Create appealing visualisations and useful application parts via the extensive library of integrated widgets and approved add-ons.
- Easily create and white label products utilising a comprehensive set of APIs and developer toolkits.
- Differentiate goods and services with fully integrated, white-label analytics in the workplace or for people travelling by mobile devices.

Sisense is a business analytics tool that offers interactive visual analytics, cleans your data and provides insights.

Sisense offers dashboards that are specifically designed for your industry. The system can gather information from various sources, offering a level of analysis to aid in decision-making and finally producing the visual analytics via dashboards that your reporting requires.

5.3.7 Grafana

Through Grafana's dashboards, information gathered from many sources is given a deeper level of meaning. Following that, these dashboards can be made available to other teams and team members, fostering cooperation and allowing for a more in-depth analysis of the data and its effects. Create dashboards specifically for you and your team using extensive querying and transformation features and alter your panels to deliver the visuals you want.

When attempting to swiftly identify the source of an incident or unusual system activity, understanding all pertinent data and data linkages is essential. Grafana enables seamless data visualisation and transfer between teams and team members, enabling quick identification and resolution of issues.

Insights from Grafana dashboards can be shared:

- Across a company—even to coworkers who aren't Grafana users themselves.
- Across the Grafana community, anywhere in the world
- Wherever you go: you can see your dashboards on all your devices wherever you are.

An observability platform for visualising metrics, logs and traces gathered from your applications is called Grafana. It's a cloud-native method for rapidly putting together data dashboards that allow you to look at and evaluate your stack.

Prometheus, InfluxDB, ElasticSearch and conventional relational database engines are just a few of the data sources that Grafana may connect to. With the aid of these

Notes

sources, you can choose the pertinent fields from your data to build intricate dashboards. Graphs, heat maps and histograms are just a few examples of the various visualisation elements that dashboards might include.

When Should Grafana Be Used?

The most popular usage of Grafana is as an infrastructure monitoring tool for tracking application performance and error rates. Visual dashboards give you real-time insights without requiring you to manually sort through data points, making it quick and simple to determine whether your stack is functioning regularly.

Grafana is excellent when people need to quickly ingest enormous amounts of raw data. Grafana gives you a single purpose-built platform that ranges from overview dashboards to advanced source interrogation, whereas your other tools, like Prometheus, may already have some level of data analysis support.

When you need to access several sources at once, it is another important use case. When it comes to finding time series events, log entries and custom queries side by side for immediate consumption, Grafana excels.

On a single page, you might create an overview dashboard that displays the hardware resource usage, significant log lines and a graph of new user sign-ups in your database. This would provide you with a single location to go to when you need an overview of everything happening in your company.

Data from all the sources that are pertinent to you may be combined, analysed and visualised using Grafana, a data analytics tool. It contains built-in support for more than 15 widely used databases and monitoring programmes. Your data sources' measurements are shown as modular panels in user-friendly dashboards to produce perspectives that everyone can comprehend.

Using Grafana for analytics helps solve several problems with data-driven DevOps. When data is inconsistent, dispersed across numerous platforms, or too sophisticated for team members who aren't data specialists to query, it frequently remains underutilised. Grafana combines all of your data into a single platform and provides you with the resources to investigate events and create practical visualisations.

Grafana has gained popularity since it enables you to leverage your data. Teams and organisations that effectively use data are better able to identify trends, make focused adjustments and increase their overall performance. Product managers, data analysts and engineers may access shared views through Grafana dashboards, which helps to keep everyone on the same page.

5.4 Case Study

5.4.1 Case study

Heathrow Airport in London acts as a global entry point. It is the second-busiest international airport in the world after Dubai International Airport. Also, it is the eighth largest in terms of total passenger throughput.

The Challenge

Given that the airport is the seventh busiest in the world in terms of overall passenger flow, one can only imagine the level of effectiveness and efforts expected from

Notes

the ground management. Airport managers and ground staff may have a challenging task managing more than 2,00000 passengers every day.

All departments must function flawlessly in unison and collaboration for the airport to efficiently handle passenger traffic and give travellers a great airport experience. At such busy airports, there are fresh challenges and uncertainties every day. Everything was disturbed by unanticipated hitches in the airport's normally smooth flow of operations.

Stormy weather rescheduled or postponed flights, modifications to the jet stream and other variables may cause issues and obstruct an airport's smooth functioning. These problems create turmoil for both airport employees and visitors.

The airport needed a centralised digital management system to handle this problem. Such a system would make use of the enormous amounts of data being produced by the operating systems of the airport and transform it into informative visual data. The interpretations produced by the BI tool can be used by airport staff to enhance operation and passenger management.

The Change

The Heathrow Group chose Microsoft Power BI for corporate intelligence and Microsoft Azure for cloud services. Microsoft Azure technology has been used to collect data from the airport's operational back-end systems. Examples of these systems include check-in counters, luggage tracking systems, airline timetables, weather tracking systems, freight tracking and numerous others. These gadgets give business intelligence applications like Power BI operational data. In Power BI, users can turn this data into useful information that airport staff can use.

Power BI organises the disorganised data into clear visualisations that show various statistics and statuses of the airport system. The ground staff, which includes baggage handlers, gate agents and air traffic controllers, then uses this information to operate and manage passengers in an effective manner.

Services like Azure Stream Analytics, Azure Data Lake Analytics and Azure SQL Database are used to extract, clean and prepare operational data in real-time. This information includes information on passenger transfers, security queues, flight movements and immigration lines. Ultimately, Power BI interprets and analyses data from a variety of Azure services.

Operational data for Power BI comes from various data sources. Using Power BI capabilities, like as visual dashboards, reports and visualisations, the data is then translated into actionable insights. About 75,000 airport employees now have information at their fingertips thanks to Power BI.

To further understand this, let's take a relevant example from the real world. Up to 20 aircraft could experience a daily delay as a result of a disruption in the jet stream. As a result, there will always be about 6,000 passengers waiting at the airport. As a result, both the volume and density of airport passengers will increase. Power BI performs similar duties to a centralised information system.

The airport uses it to inform visitors of the dramatic spike in traffic. Many airport departments, including food outlets, immigration, customs, gate attendants and luggage handlers, receive this information. They will have adequate time to prepare to help the travellers because of this.

Airport staff may be informed in advance of anticipated delays and unforeseen spikes in passenger traffic thanks to smart BI solutions like Power BI. In order to prevent

Notes

any last-minute hiccups, this enables management teams and other staff members to take necessary measures in advance, such as increasing food stock levels, adding more passenger buses, increasing ground staff levels, directing passengers to the waiting area, etc.

Heathrow has reaped various benefits from utilising a powerful BI tool like Power BI. They are ecstatic that Power BI's capabilities have allowed them to give their clients hassle-free airport experiences. Heathrow is also extending Power BI apps in order to forecast passenger movement at the airport and prevent any unexpected issues for the passengers.

Summary

- An open-source web tool called the Jupyter Notebook enables you to create and share documents with real-time code, equations, visuals and text.
- ModelOps will allow you to oversee and manage the AI lifecycle, while prescriptive analytics will help you make better business decisions and visual modelling tools will speed up time to value
- Python is a popular computer programming language used to create software and websites, automate processes and analyse data. Python is a general-purpose language, which means it can be used to make many various types of programmes and isn't tailored for any issues.
- For software developers, GitHub is a web-based platform for version control and collaboration. The largest single donor to GitHub, Microsoft, purchased the service for $7.5 billion in 2018.
- On Power BI Desktop, you can build a variety of reports. The Power BI service allows these reports to be published on the Power BI dashboard. By selecting the Publish button, a Power BI report prepared on Power BI Desktop can be uploaded to Power BI Service.
- The robust and rapidly expanding data visualisation tool is Tableau. Tableau is a business intelligence platform that enables us to evaluate raw data visually in the form of graphs, reports, etc.
- A tool for data visualisation called Infogram enables you to make interactive maps, infographics and charts. All throughout the world, newsrooms, marketing teams and students use it.

Glossary

- Package: A package is a means to ensure that none of the names you select to use in your programme "step on the toes" of names that another programme might use.
- Kernal: The language-specific processing tools used to process notebook cells are called notebook kernels.
- Power Query: The data transformation and mashup engine are called Power Query.
- Power Pivot: A data modelling tool called Power Pivot enables you to build data models, establish associations and do computations.

Check your Understanding

1. Which language is required while installing Jupyter Notebook?

   a) C++ b) SQL

   c) Python d) Java

Notes

2. What factors should be considered while selecting software?
   a) Customization b) Requirement
   c) Value d) All the above
3. Which is/are the essential part of Jupyter notebook?
   a) Cell b) A runtime environment
   c) A file system d) All the above
4. In which mode of Jupyter notebook you can move between cells, add and remove cells and modify the cell type?
   a) Edit mode b) Command mode
   c) Read Mode d) Write mode
5. What is the option to build a list of bulleted items in Jupyter Notebook workspace?
   a) begin successive lines with a
   b) start it with a number, then add a period
   c) enclose it in **, __, or ‘
   d) start it with a number, then add a period
6. In how many hours session will be closed automatically in Jupyter lab?
   a) 6 b) 8
   c) 12 d) 24
7. Which Kernal is required for query service?
   a) Scala b) R
   c) SQL d) Python
8. How many windows will appear when initially launch Rstudio?
   a) 3 b) 4
   c) 5 d) 2
9. What is advantage of IBM Watson?
   a) It does not immediately process structured data
   b) it can learn more from less
   c) Despite the volume of data increasing, there are still little resources available to meet the demands
   d) All the above
10. What is barrier of using IBM Watson?
   a) High Switching Cost
   b) it limits the areas of use
   c) It does not immediately process structured data
   d) maintenance is a significant concern
11. What is the data model of MySQL?
   a) Column-Family b) Document
   c) Keyvalue d) Relational

Notes

12. The software that runs self-driving cars and Netflix's recommendation engine were both developed using ____.
    a) R b) Java
    c) Python d) SQL
13. What is/are the thing/s are done with python?
    a) Data analysis and machine learning b) Web development
    c) Automation or scripting d) All the above
14. Which library Python provides that help programmers to create data analysis and machine learning applications more rapidly and effectively?
    a) TensorFlow b) Knime
    c) R d) None of the above
15. In which year GitHub was established?
    a) 2009 b) 2008
    c) 2007 d) 2006
16. What is the advantage of Power BI?
    a) Limited data processing capabilities
    b) Limited customization options
    c) Customizable dashboards
    d) Users who want more features or storage space from Power BI may have to pay for it since it is not a free service
17. Who is/are the user/s of Power BI?
    a) Business analysts b) Business intelligence analysts
    c) Supply chain analysts d) All the above
18. A full data discovery and analytics platform is ____.
    a) Tableau b) Qlinksense
    c) PowerBI d) GitHub
19. What is/are the Zoho analytics pros?
    a) Gives an immediate snapshot into many KPIs.
    b) Enables you to analyse your data in real time.
    c) Allows you to create and share visual dashboards.
    d) All the above
20. ____, whose name is derived from the Japanese word for "Thank You," is a cloud-based solution created to combine data from various back-end systems to provide comprehensive views of organisational performance with filterable and customizable access to a range of key performance indicators (KPIs) and metrics.
    a) Domo b) Zoho
    c) Knime d) Keras

Exercise

1. Explain Jupyter notebook.

Notes

2. Analyse IBM Watson studio.
3. Describe Jupyter Labs
4. What is the different IBM tool?
5. What is different type of visualisation package?

Learning Activities

1. Explain Tableau with example.
2. What is Qlik sense explain with example?

Check your Understanding-Answers

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

Further Readings and Bibliography

1. Field Cady. The Data Science. 2017
2. William Vance. Data Science: 3 Book in 1 – Beginner's Guide to learn the Realm of Data Science. 2020
3. Peter Bruce Andrew Bruce. Practical Statistics for Data Scientist. 2020
4. Reema Thareja. Data Science and Machine Learning using Python. 2022
5. Uma Maheshwari R Sujatha. Introduction to Data Science: Practical Approach with R and Python. 2021