MRCET CAMPUS

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

(AUTONOMOUS INSTITUTION – UGC, GOVT. OF INDIA)

Department of CSE
(Emerging Technologies)
(Data Science)

B.TECH(R-22)
(II YEAR – I SEM)
(2023-24)

Data Science & its Applications
(R22A6701)

LECTURE NOTES

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous Institution – UGC, Govt. of India)
Recognized under 2(f) and 12(B) of UGC ACT 1956
(Affiliated to JNTUH, Hyderabad, Approved by AICTE-Accredited by NBA & NAAC – 'A' Grade - ISO 9001:2015 Certified)
Maisammaguda, Dhulapally (Post Via. Hakimpet), Secunderabad–500100, Telangana State, India

Department of Computer Science and Engineering

EMERGING TECHNOLOGIES

Data Science & Its Applications
(R22A6701)

LECTURE NOTES

Prepared by
P.Sreenivas, Associate Professor & G.Gayatri, Assistant Professor

On
4.08.2023

Department of Computer Science and Engineering

EMERGING TECHNOLOGIES

Vision

- "To be at the forefront of Emerging Technologies and to evolve as a Centre of Excellence in Research, Learning and Consultancy to foster the students into globally competent professionals useful to the Society."

Mission

The department of CSE (Emerging Technologies) is committed to:

- To offer highest Professional and Academic Standards in terms of Personal growth and satisfaction.
- Make the society as the hub of emerging technologies and thereby capture opportunities in new age technologies.
- To create a benchmark in the areas of Research, Education and Public Outreach.
- To provide students a platform where independent learning and scientific study are encouraged with emphasis on latest engineering techniques.

QUALITY POLICY

- To pursue continual improvement of teaching learning process of Undergraduate and Post Graduate programs in Engineering & Management vigorously.
- To provide state of art infrastructure and expertise to impart the quality education and research environment to students for a complete learning experiences.
- Developing students with a disciplined and integrated personality.
- To offer quality relevant and cost effective programmes to produce engineers as per requirements of the industry need.

For more information: www.mrcet.ac.in

SYLLABUS

I. *COURSE OBJECTIVES:*

The students will try to learn:

I. The fundamental knowledge on basics of data science.
II. The programs in R language for understanding and data manipulation using R
III. The fundamentals of how to obtain, store, explore, and model data efficiently.
IV. The knowledge on Data Science Application and its Tools.

Unit-I: Introduction to Data Science- Introduction- Definition - Data Science in various fields - Examples - Impact of Data Science - Data Analytics Life Cycle - Data Science Toolkit - Data Scientist - Data Science Team Understanding data: Introduction – Types of Data: Numeric –Categorical – Graphical – High Dimensional Data – Classification of digital Data: Structured, Semi-Structured and UnStructured - Example Applications. Sources of Data: Time Series – Transactional Data – Biological Data – Spatial Data – Social Network Data – Data Evolution. (From AU)

Unit-II: R Programming: Introduction to R- Features of R - Environment - R Studio. Basics of R-Assignment - Modes - Operators - special numbers - Logical values - Basic Functions - R help functions - R Data Structures - Control Structures. Vectors: Definition- Declaration - Generating - Indexing
- Naming - Adding & Removing elements - Operations on Vectors - Recycling
- Special Operators - Vectorized if- then else-Vector Equality – Functions for vectors
- Missing values - NULL values - Filtering & Subsetting. (From AU)

Unit-III Exploratory Data Analysis and the Data Science Process - Exploratory Data Analysis and the Data Science Process - Basic tools (plots, graphs and summary statistics) of EDA - Philosophy of EDA - The Data Science Process - Case Study: Data collection process in real time applications.

Unit-IV Data Science Applications - Data Science and it's various applications – Data Science Applications in Uses Cases Applications of Data Science - In Search Engines, Social Media, Transportation, Banking, Financial Services and Insurance (BFSI), Business and E-Commerce & Retail Applications, Health Care Sector, Targeting Recommendation, Gaming Technology, Medicine and Drug Development and Telecom etc.Introduction- Collecting and Analyzing Twitter Data and YouTube Data.

Unit-V: Data Science Toolkit: Brief Introduction to data science tools: SaS, Apache Spark, BigML, Excel, R-Programming, TensorFlow, KNIME, Tableau, PowerBI etc with advantages and disadvantages.

III. *TEXT BOOKS:*

1. Sinan Ozdemir, “Principles of Data Science”, Packt.
2. Norman Matloff , “The Art of R Programming”, Cengage Learning.

IV. *REFERENCE BOOKS:*

1. Cathy O’Neil and Rachel Schutt, “Doing Data Science, Straight Talk From The Frontline”, O’Reilly, 2014.
2. Nina Zumel, John Mount, “Practical Data Science with R”, Manning Publications, 1st Edition, 2014.
3. 3. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly,2015.

V. *WEB REFERENCES:*

1. https://en.wikipedia.org/wiki/R_programming_language
2. http://www.r-bloggers.com/how-to-learn-r-2/#h.obx6jyuc9j7t.
3. http://www.tutorialspoint.com/r/

VI. *E BOOKS*

1. https://www.programmer-books.com/introducing-data-science-pdf/
2. https://www.cs.uky.edu/~keen/115/Haltermanpythonbook.pdf
3. https://innovacion-tecnologia.com/wp-content/uploads/2020/09/DATA-SCIENCE-FROM-SCRATCH.pdf
4. https://covid19.uthm.edu.my/wp-content/uploads/2020/04/Data-Science-from-Scratch-First-Principles-with-Python-by-Joel-Grus-z- lib.org_.epub_.pdf

VII. *URSE OUTCOMES:*

1. Describe what Data Science is and the skill sets needed to be a datascientist
2. Ability to learn the R Programming
3. Explain the significance of exploratory data analysis (EDA) in data science.
4. Explore the Various Data Science Applications
5. Understand the various tools for Data Science and its Analysis

INDEX

Definition :

Data science involves using methods to analyze massive amounts of data and extract the knowledge it contains. There Relationship between bigdata and data science as being like the relationship between crude oil and an oil refinery. Data science is an evolutionary extension of statistics capable of dealing with the massive amounts of data produced.

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into tangible business value.

Data Science in various fields :

Data Science provides multiple application , with deep study of a large quantity of data, which involves extracting some meaningful from the raw, structured, and unstructured data. The extracting out meaningful data from large amounts use processing of data and this processing can be done using statistical techniques and algorithm, scientific techniques, different technologies, etc.

Image Recognition : Data Science is used in Image Recognition. When we upload image with our friend on Facebook, Facebook gives suggestions Tagging who is in the picture. This is done with the help of machine learning and Data Science. When an Image is Recognized, the data analysis is done on one's Facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile then Facebook suggests us auto-tagging.

Airline Routing Planning : Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or take a halt in between like a flight can have a direct route from Delhi to the U.S.A or it can halt in between after that reach at the destination.

Medicine and Drug Development :creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone's life. Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors.

Finance :Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company. Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.

Impact of Data Science :

Quantifiable & Data-Driven Decision Making

This is arguably the biggest reason many businesses utilize data science applications, and its usually also the biggest benefit One relatively new but exciting feature of this technology is the ability to analyze streaming data through time series analysis, giving businesses real-time feedback that they can act on.

Recruiting

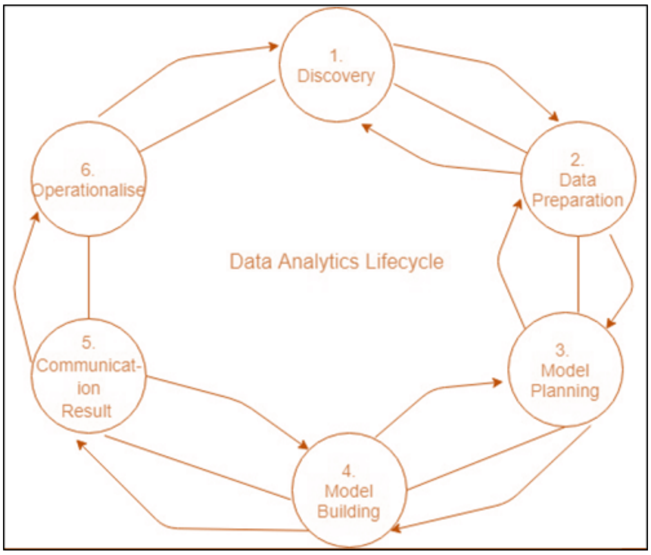
Recruiting and retaining quality and skilled employees is a struggle for many businesses, regardless recruiting by automating aspects of the recruiting process to help organizations find better candidates, faster.

Opportunity Identification

Another capability of data science tools and analytics is opportunity identification. Using historical and forecasted market data, businesses can identify geographic areas to target to penetrate for sales and marketing initiatives with greater accuracy.

Data Analytics Life Cycle :

The Data analytics lifecycle was designed to address Big Data problems and data science projects. The process is repeated to show the real projects. To address the specific demands for conducting analysis on Big Data, the step-by-step methodology is required to plan the various tasks associated with the acquisition, processing, analysis, and recycling of data.



1. Data Discovery :

This is the initial phase to set your project's objectives and find ways to achieve a complete data analytics lifecycle. Start with defining your business domain and ensure you have enough resources (time, technology, data, and people) to achieve your goals.

The biggest challenge in this phase is to accumulate enough information, need to draft an analytic plan, which requires some serious leg work.

Accumulate resources

To analyse the models you have intended to develop. Then determine how much domain knowledge you need to acquire for fulfilling those models.

2. Data Preparation and Processing :

The Data preparation and processing phase involves collecting, processing, and conditioning data before moving to the model building process.

Identify data sources

Data have to identify various data sources and analyse how much and what kind of data you can accumulate within a given timeframe. Evaluate the data structures, explore their attributes and acquire all the tools needed.

Collection of data

can collect data using three methods:

Data acquisition: can collect data through external sources.

Data Entry: can prepare data points through digital systems or manual entry as well.

Signal reception: can accumulate data from digital devices such as IoT devices and control systems.

3. Model Planning :

This is a phase where analysist has to analyse the quality of data and find a suitable model for the project.

Loading Data in Analytics Sandbox

A data sandbox is a tool that allows to test data in a safe environment without affecting the actual data. It will enable to play around with different ways of using data and see what happens without causing any damage or danger to existing data.

Analytics Sandbox : An analytics sandbox is a part of data lake architecture that allows to store and process large amounts of data. It can efficiently process a large range of data such as big data, transactional data, social media data, web data, and many more.

Data are loaded in the sandbox in three ways:

ETL –(Extract, Transform And Load) : Team specialists make the data comply with the business rules before loading it in the sandbox.

ELT − The data is loaded in the sandbox and then transform as per business rules.

ETLT − It comprises two levels of data transformation, including ETL and ELT both.

The data have collected may contain unnecessary features or null values. It may come in a form too complex to anticipate. This is where data exploration' can help uncover the hidden trends in data.

Steps involved in data exploration:

- Data identification
- Univariate Analysis
- Multivariate Analysis
- Filling Null values

For model planning, data analysts often use regression techniques, decision trees, neural networks, etc. Tools mostly used for model planning and execution include Rand PL/R, WEKA, Octave, Statista, and MATLAB.

4. Model Building :

Model building is the process where you have to deploy the planned model in a real-time environment. It allows analysts to solidify their decision-making process by gain in-depth analytical information. This is a repetitive process, as you have to add new features as required by your customers constantly.

In some cases, a specific model perfectly aligns with the business objectives data, and sometimes it requires more than one try.

5. Communication Result :

This is the phase where you have to communicate the data analysis with your clients. It requires several intricate processes where how to present information to clients in a lucid manner. clients don't have enough time to determine which data is essential. Therefore, must do an impeccable job to grab the attention of your clients.

Check the data accuracy

Is the data provide information as expected? If not, then you have to run some other processes to resolve this issue. need to ensure the data process provides consistent information. This will help to build a convincing argument while summarizing your findings.

Highlight important findings

Well, each data holds a significant role in building an efficient project. However, some data inherits more potent information that can truly serve your audience's benefits. While summarizing your findings, try to categorize data into different key points.

Determine the most appropriate communication format

How to communicate findings tells a lot about as a professional. We recommend you to go for visuals presentation and animations as it helps to convey information much faster. However, sometimes also need to go old-school as well. For instance, clients may have to carry the findings in physical format. They may also have to pick up certain information and share them with others.

6. Operationalize :

As soon prepare a detailed report including key findings, documents, and briefings, data analytics life cycle almost comes close to the end. The next step remains the measure the effectiveness of analysis before submitting the final reports to stakeholders.

In this process, have to move the sandbox data and run it in a live environment. Then you have to closely monitor the results, ensuring they match with expected goals. If the findings fit perfectly with objective, then can finalize the report. Otherwise, have to take a step back in data analytics lifecycle and make some changes.

Data Science Toolkit :

Data Science Toolkit in Data Science is the art of drawing and visualizing useful insights from data. it is the process of collecting, analyzing, and modeling data to solve problems related to the real-world. To implement the operations we have to use such tools to manipulate the data and entities to solve the issues. There are pre-defined functions, algorithms, and a user-friendly Graphical User Interface (GUI). As we know that Data Science has a very fast execution process, one tool is not enough to implement this.

Most Frequent Used Tools For Data Science :

1. Apache Hadoop

Apache Hadoop is a free, open-source framework by **Apache Software Foundation** authorized under the Apache License 2.0 that can manage and store tons and tons of data. It is used for high-level computations and data processing. By using its parallel processing nature, we can work with the number of clusters of nodes.

- Hadoop offers standard libraries and functions for the subsystems.
- Effectively scale large data on thousands of Hadoop clusters.
- It speeds up disk-powered performance by up to 10 times per project.

2. SAS (Statistical Analysis System)

SAS is a statistical tool developed by **SAS Institute.** It is a closed source proprietary software that is used by large organizations to analyze data. It is one of the oldest tools developed for Data Science. It is used in areas like *Data Mining, Statistical Analysis, Business Intelligence Applications, Clinical Trial Analysis, Econometrics & Time-Series Analysis.*

Latest Version: SAS 9.4

- It is a suite of well-defined tools.
- It has a simple but most effective GUI.

- It provides a Granular analysis of textual content.

3. Apache Spark

Apache Spark is the data science tool developed by **Apache Software Foundation** used for analyzing and working on large-scale data. It is a unified analytics engine for large-scale data processing. It is specially designed to handle batch processing and stream processing.

Latest Version: Apache Spark 2.4.5

- It offers data cleansing, transformation, model building & evaluation.
- It has the ability to work in-memory makes it extremely fast for processing data and writing to disk.
- It provides many APIs that facilitate repeated access to data.

4. Data Robot

DataRobot Founded in 2012, is the leader in enterprise AI, that aids in developing accurate predictive models for the real-world problems of any organization. It facilitates the environment to automate the end-to-end process of building, deploying, and maintaining your AI. DataRobot's Prediction Explanations help you understand the reasons behind your machine learning model results.

- Highly Interpretable.
- It has the ability to making the model's predictions easy to explain to anyone.
- It provides the suitability to implement the whole Data Science process at a large scale.

5. Tableau

Tableau is the most popular data visualization tool used in the market, is an American interactive data visualization software company founded in January 2003, was recently acquired by Salesforce.

Latest Version: Tableau 2020.2

- It offers comprehensive end-to-end analytics.
- It is a fully protected system that reduces security risks to the maximum state.
- It provides a responsive user interface that fits all types of devices and screen dimensions.

6. BigML

BigML, founded in 2011, is a Data Science tool that provides a fully interactable, cloud-based GUI environment that you can use for processing Complex Machine Learning Algorithms. The main goal of using BigML is to make building and sharing datasets and models easier for everyone. It provides an environment with just one framework for reduced dependencies.

Latest Version: BigML Winter 2020

- It specializes in predictive modeling.
- It has ability to export models via JSON PML and PMML makes for a seamless transition from one platform to another.
- It provides an easy to use web-interface using Rest APIs.

7. TensorFlow

TensorFlow, developed by **Google Brain team**, is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It provides an environment for building and training models, deploying platforms such as computers, smartphones, and servers, to achieving maximum potential with finite resources. It is one of the very useful tools that is used in the fields of Artificial Intelligence, Deep Learning, & Machine Learning.

Latest Version: TensorFlow 2.2.0

- It provides good performance and high computational abilities.
- Can run on both CPUs and GPUs.
- It provides features like easily trainable and responsive construct.

8. Jupyter

Jupyter, developed by **Project Jupyter** on February 2015 open-source software, open-standards, and services for interactive computing across dozens of programming languages. It is a web-based application tool running on the kernel, used for writing live code, visualizations, and presentations

Latest Version: Jupyter Notebook 6.0.3

- It provides an environment to perform data cleaning, statistical computation, visualization and create predictive machine learning models.

Data Scientist :

Data Scientist Skills : A good data scientist will have the right combination of technical and non-technical skills in their toolkit. We have compiled a list of important skills that enhance the job of data scientist.

Technical Skills:

- Programming languages like Structured Query Language (SQL), Python, Statistical Analysis System (SAS), etc
- Machine learning (ML) and deep learning
- Data visualization tools such as Tableau, PowerBI, etc
- Statistical analysis
- Data wrangling

Non-Technical Skills:

- Effective communication
- Proactive problem-solving
- Strong business acumen
- Solid critical and analytical thinking

Data Scientist Roles and Responsibilities

The job of a data scientist goes beyond interpreting large data sets to derive actionable insights. These are the main roles and responsibilities of a data scientist:

- Extract and mine relevant data sources that match business needs
- Collect both structured and unstructured data sets to perform data analysis
- Develop ML algorithms and prediction systems
- Use ML tools to improve data quality

- Interpret data and ensure data uniformity to identify useful trends and patterns

Data Science Careers

A career in data science is highly lucrative as the demand for professionals in this cutting-edge field is rising remarkably. However, the role of a data scientist is not the only job to pursue or aim for in this field. In fact, data science offers a diverse range of careers that are actually shaping the future. Here are some of the most popular roles in data science.

- Machine learning engineer
- Data analyst
- Applications architect
- Business intelligence (BI) developer
- Marketing analyst
- Database administrator

Types of Data:

Qualitative (Categorical) Data :

Qualitative data usually describes an object or a group of items. It's also known as categorical data because, as the name implies, you can label a group of items or data points to a specific category. Examples include colors, plants, and places.

Qualitative data is then classified into 2 other subtypes – "ordinal" and "nominal".

Ordinal Data :

Ordinal data follows a specific order or ranking, as in test grades, economic status, or military rank.

Nominal Data :

Nominal data, however, doesn't follow a specific order like ordinal data. Consider gender, city, employment status, colors, etc.

Quantitative (Numerical) Data :

On the other hand, quantitative data deals with numeric values on which we can apply mathematical operations – height, fruits in a basket, kids in a school.

Although they seem similar, here's something else to keep in mind – quantitative data can be continuous or discrete.

The difference is that we can split continuous data further into smaller units, and they still make sense. However, this is not possible with discrete data, as dividing them into smaller units will give us unreasonable values.

For example, weight is continuous because we can measure it in kilograms, grams, and milligrams and still we have a valid weight value.

Frequency tables, pie charts, and bar charts are the most appropriate graphical displays for categorical variables. Below are a frequency table, a pie chart, and a bar graph for data concerning Mental Health Admission numbers.
High-dimensional data are defined as data in which the number of features (variables observed), p, are close to or larger than the number of observations (or data points), n.

Classification of digital Data: Structured, Semi-Structured and UnStructured :

Digital data can be classified into three forms:

1. Unstructured Data
2. Semi-Structured Data
3. Structured

Structured Data

- Able to be processed, sorted, analyzed, and stored in a predetermined format, then retrieved in a fixed format
- Accessed by a computer with the help of search algorithms
- First type of big data to be gathered

- Easiest of the three types of big data to analyze
- Examples of structured data include:
    - Application-generated data
    - Dates
    - Names
    - Numbers (e.g., telephone, credit card, US ZIP Codes, social security)

Semi-Structured Data

- Contains both structured as well as unstructured information
- Data may be formatted in segments
- Appears to be fully-structured, but may not be
- Not in the standardized database format as structured data
- Has some properties that make it easier to process than unstructured data
- Examples
    - CSV
    - Electronic data interchange (EDI)
    - HTML
    - JSON documents
    - NoSQL databases
    - Portable Document Files (PDF)
    - RDF
    - XML

Unstructured Data

- Not in any predetermined format (i.e., no apparent format)
- Accounts for the majority of the digital data that makes up big data
- Examples of the different types of unstructured data include:
    - Human-generated data
        - Email
        - Text messages
        - Invoices

  - Text files
  - Social media data
- Machine-generated data
  - Geospatial data
  - Weather data
  - Data from IoT and smart devices
  - Radar data
  - Videos
  - Satellite images
  - Scientific data

Time Series :

A time series is a sequence of observations measured at succesive times. Time series are monthly, trimestrial, or annual, sometimes weekly, daily, or hourly (study of road traffic, telephone traffic), or biennial or decennial.

Time series analysis consists of methods that attempt to understand such time series to make predictions.

Time series can be decomposed into four components, each expressing a particular aspect of the movement of the values of the time series.

These four components are:

- **Secular trend**, which describe the movement along the term;
- **Seasonal variations**, which represent seasonal changes;
- **Cyclical fluctuations**, which correspond to periodical but not seasonal variations;
- **Irregular variations**, which are other nonrandom sources of variations of series.

The analysis of time series consists in making mathematical descriptions of these elements, that is, estimating separately the four components

UNIT-2

R-Programming

1. Introduction to R

R Programming Tutorial is designed for both beginners and professionals. Our tutorial provides all the basic and advanced concepts of data analysis and visualization.

R is a software environment which is used to analyze statistical information and graphical representation. R allows us to do modular programming using functions.

Our R tutorial includes all topics of R such as introduction, features, installation, rstudio ide, variables, datatypes, operators, if statement, vector, data handing, graphics, statistical modelling, etc. This programming language was named R, based on the first name letter of the two authors (Robert Gentleman and Ross Ihaka).

2. Features of R - Environment

- Comprehensive Language.
- Provides a Wide Array of Packages.
- Possesses a Number of Graphical Libraries.
- Open-source.
- Cross-Platform Compatibility.
- Facilities for Various Industries.
- No Need for a Compiler.
- Performs Fast Calculations

3. R Studio.

R Studio is an integrated development environment(IDE) for R. IDE is a GUI, where you can write your quotes, see the results and also see the variables that are generated during the course of programming.

- R Studio is available as both Open source and Commercial software.
- R Studio is also available as both Desktop and Server versions.
- R Studio is also available for various platforms such as Windows, Linux, and macOS.

Rstudio is an open-source tool that provides Ide to use R language, and enterprise-ready professional software for data science teams to develop share the work with their team.

4. Basics of R-Assignment

The use of these operators is to assign values to the variables. There are two kinds of assignments, leftwards assignment, and rightwards assignment. Operators '<-' and '=' are used to assign values to any variable. x<- 3 or x = 3 (Leftwards Assignment)

- Arithmetic operators
- Assignment operators
- Comparison operators
- Logical operators
- Miscellaneous operators

- numeric - (10.5, 55, 787)
- integer - (1L, 55L, 100L, where the letter "L" declares this as an integer)
- complex - (9 + 3i, where "i" is the imaginary part)
- character (a.k.a. string) - ("k", "R is exciting", "FALSE", "11.5")
- logical (a.k.a. boolean) - (TRUE or FALSE)

A variable provides us with named storage that our programs can manipulate. A variable in R can store an atomic vector, group of atomic vectors or a combination of many Robjects. A valid variable name consists of letters, numbers and the dot or underline characters.

Operators

- Arithmetic Operators
- Relational Operators
- Logical Operators
- Assignment Operators
- Miscellaneous Operators

Basic Functions

A function is a set of statements organized together to perform a specific task. R has a large number of in-built functions and the user can create their own functions.

In R, a function is an object so the R interpreter is able to pass control to the function, along with arguments that may be necessary for the function to accomplish the actions.

The function in turn performs its task and returns control to the interpreter as well as any result which may be stored in other objects.

Control Structures

Loop control statements change execution from its normal sequence. When execution leaves a scope, all automatic objects that were created in that scope are destroyed.

Vectors are the most basic R data objects and there are six types of atomic vectors. They are logical, integer, double, complex, character and raw.

UNIT-3
Exploratory Data Analysis (EDA)

Importance of EDA in Data Science

Data analysis is a broad term involving different types of analysis like descriptive, diagnostic, predictive, and prescriptive. EDA is synonymous with descriptive analysis, where one explores the hidden relationships and patterns in the available data.

Exploratory Data Analysis is important for any business. It lets data scientists analyze the data before reaching any conclusion. Also, this makes sure that the results which are out are valid and applicable to business outcomes and goals. The Data Science field is now very important in the business world as it provides many opportunities to make vital business decisions by analyzing hugely gathered data. Understanding the data thoroughly needs its exploration from every aspect. The impactful features enable making meaningful and beneficial decisions; therefore, EDA occupies an invaluable place in Data science. Let's suppose we want to make a data science project on the employee churn rate of a company. But before we make a model on this data we have to analyze all the information which is present across the dataset like as what is the salary distribution of employees, what is the bonus they are getting, what is their starting time, and the assigned team. These all steps of analyzing and modifying the data come under EDA.

Definition

Exploratory Data Analysis (EDA) is one of the techniques used for extracting vital features and trends used by machine learning and deep learning models in Data Science. That is an approach used to analyze the data and discover trends, patterns, or check assumptions in data with the help of statistical summaries and graphical representations. Thus, EDA has become an important milestone for anyone working in data science

The main underlying principles of an EDA are-

- The aim should be to uncover information that should lead to showing patterns and trends. Identifying trends in time and space
-
- Missing values and outliers need to be given proper consideration
- The relationship between different variables must be established.
- A suitable technique of variate analysis should be chosen for the target to be achieved.

Role of EDA in Data Science

The role of data exploration analysis is based on the use of objectives achieved as above. After formatting the data, the performed analysis indicates patterns and trends that help to take the proper actions required to meet the expected goals of the business. As we expect specific tasks to be done by any executive in a particular job position, it is expected that proper EDA will fully provide answers to queries related to a particular business decision. As data science involves building models for prediction, they require optimum data features to be considered by the model. Thus, EDA ensures that the correct ingredients in patterns and trends are made available for training the model to achieve the correct outcome, like a successful recipe. Therefore, carrying out the right EDA with the correct tool based on befitting data will help achieve the expected g **Steps Involved in Exploratory Data Analysis (EDA)**

The key components in an EDA are the main steps undertaken to perform the EDA. These are as follows:

1. Data Collection

Nowadays, data is generated in huge volumes and various forms belonging to every sector of human life, like healthcare, sports, manufacturing, tourism, and so on. Every business knows the importance of using data beneficially by properly analyzing it. However, this depends on collecting the required data from various sources through surveys, social media, and customer reviews, to name a few. Without collecting sufficient and relevant data, further activities cannot begin.

2. Finding all Variables and Understanding Them

When the analysis process starts, the first focus is on the available data that gives a lot of information. This information contains changing values about various features or characteristics, which helps to understand and get valuable insights from them. It requires first identifying the important variables which affect the outcome and their possible impact. This step is crucial for the final result expected from any analysis.

3. Cleaning the Dataset

The next step is to clean the data set, which may contain null values and irrelevant information. These are to be removed so that data contains only those values that are relevant and important from the target point of view. This will not only reduce time but also reduces the computational power from an estimation point of view. Preprocessing takes care of all issues, such as identifying null values, outliers, anomaly detection, etc.

4. Identify Correlated Variables

Finding a correlation between variables helps to know how a particular variable is related to another. The correlation matrix method gives a clear picture of how different variables correlate, which further helps in understanding vital relationships among them.

5. Choosing the Right Statistical Methods

As will be seen in later sections, depending on the data, categorical or numerical, the size, type of variables, and the purpose of analysis, different statistical tools are employed. Statistical formulae applied for numerical outputs give fair information, but graphical visuals are more appealing and easier to interpret.

6. Visualizing and Analyzing Results

Once the analysis is over, the findings are to be observed cautiously and carefully so that proper interpretation can be made. The trends in the spread of data and correlation between variables give good insights for making suitable changes in the data parameters. The data analyst should have the requisite capability to analyze and be well-versed in all analysis techniques. The results obtained will be appropriate to data of that particular domain and are suitable for use in retail, healthcare, and agriculture.

Aspiring data science professionals must understand and practice the above EDA data science steps to master exploratory data analysis.

Types of EDA

Depending on the number of columns we are analyzing we can divide EDA into three types.

1. **Univariate Analysis** – In univariate analysis, we analyze or deal with only one variable at a time. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
2. **Bi-Variate analysis –** This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship between the two variables.
3. **Multivariate Analysis** – When the data involves three or more variables, it is categorized under multivariate. e.g., type of product and quantity sold against the product price, advertising expenses, and discounts offered.

The result of the analysis can be represented in numerical values, visualization, or graphical form. Accordingly, they could be further sub categorize EDA into two parts as non-graphical or graphical.

1. Non-graphical Analysis – In non-graphical analysis, we analyze data using statistical tools like mean median or mode or skewness
2. Graphical Analysis – In graphical analysis, we use visualizations charts to visualize trends and patterns in the data

.

1. Univariate Non-Graphical

The main aim of univariate non-graphical EDA is to find out the details about the distribution of the population data and to know some specific parameters of statistics. The significant parameters which are estimated from a distribution point of view are as follows:

- **Central Tendency:** This term refers to values located at the data's central position or middle zone. The three generally estimated parameters of central tendency are mean, median, and mode. Mean is the average of all values in data, while the mode is the value that occurs the maximum number of times. The Median is the middle value with equal observations to its left and right.
- **Range:** The range is the difference between the maximum and minimum value in the data, thus indicating how much the data is away from the central value on the higher and lower side.
- **Variance and Standard Deviation:** Two more useful parameters are standard deviation and variance. Variance is a measure of dispersion that indicates the spread of all data points in a data set.
- standard deviation is the square root value of it. The larger the value of standard deviation, the farther the spread of data, while a low value indicates more values clustering near the mean.

2. Univariate Graphical

Some common types of univariate graphics are:

- **Stem-and-leaf Plots:** This is a very simple but powerful EDA method used to display quantitative data but in a shortened format. It displays the values in the data set, keeping each observation intact but separating them as stem (the leading digits) and remaining or trailing digits as leaves. But histogram is mostly used in its place now.
- **Histograms (Bar Charts):** The simplest fundamental graph is a histogram, which is a bar plot with each bar representing the frequency, i.e., the count or proportion (the ratio of count to the total count of occurrences) for various values. . Histograms are very simple to quickly understand your data, which tell about values of data like central tendency, dispersion, outliers, etc.
- These plots are used to display both grouped or ungrouped data. On the x-axis, values of variables are plotted, while on the y-axis are the number of observations or frequenciesThere are many types of histograms, a few of which are listed below:

1. **Simple Bar Charts:** These are used to represent categorical variables with rectangular bars, where the different lengths correspond to the values of the variables.

2. **Multiple or Grouped charts:** Grouped bar charts are bar charts representing multiple sets of data items for comparison where a single color is used to denote one specific series in the dataset.
3. **Percentage Bar Charts:** These are bar graphs that depict the data in the form of percentages for each observation.
4. **Box Plots:** These are used to display the distribution of quantitative value in the data. If the data set consists of categorical variables, the plots can show the comparison between them. Further, if outliers are present in the data, they can be easily identified. These graphs are very useful when comparisons are to be shown in percentages, like values in the 25 %, 50 %, and 75% range (quartiles).

3. Multivariate Non-Graphical

The multivariate non-graphical exploratory data analysis technique is usually used to show the connection between two or more variables with the help of either cross-tabulation or statistics.

- For categorical data, an extension of tabulation called cross-tabulation is extremely useful. For two variables, cross-tabulation is preferred by making a two-way table with column headings that match the amount of one variable and row headings that match the amount of the opposite two variables, then filling the counts with all subjects that share an equivalent pair of levels.
- For each categorical variable and one quantitative variable, we can generate statistical information for quantitative variables separately for every level of the specific variable. We then compare the statistics across the number of categorical variables.

4. Multivariate Graphical

Graphics are used in multivariate graphical data to show the relationships between two or more variables. Here the outcome depends on more than two variables, while the change-causing variables can also be multiple.

Some common types of multivariate graphics include:

A) Scatter Plot

The essential graphical EDA technique for two quantitative variables is the scatter plot, so one variable appears on the x-axis and the other on the y-axis and, therefore, the point for every case in your dataset. This can be used for bivariate analysis.

B) Multivariate Chart

A Multivariate chart is a type of control chart used to monitor two or more interrelated process variables. This is beneficial in situations such as process control, where engineers are likely to benefit from using multivariate charts. These charts allow monitoring multiple parameters together in a single chart. A notable advantage of using multivariate charts is that they help minimize the total number of control charts for organizational processes. Pair plots generated using the Seaborn library are a good example of multivariate charts as they help visualize the relationships between all numerical variables in the entire dataset at once.

C) Run Chart

A run chart is a data line chart drawn over time. In other words, a run chart visually illustrates the process performance or data values in a time sequence.. A trend chart or time series plot is another name for a run chart.

D) Bubble Chart

Bubble charts scatter plots that display multiple circles (bubbles) in a two-dimensional plot. These are used to assess the relationships between three or more numeric variables. In a bubble chart, every single dot corresponds to one data point, and the values of the variables for each point are indicated by different positions such as horizontal, vertical, dot size, and dot colors.

E) Heat Map

A heat map is a colored graphical representation of multivariate data structured as a matrix of columns and rows. The heat map transforms the correlation matrix into color coding and represents these coefficients to visualize the strength of correlation among variables. It assists in finding the best features suitable for building accurate Machine Learning models.

Apart from the above, there is also the ‘Classification or Clustering analysis’ technique used in EDA. It is an unsupervised type of machine learning used for the classification of input data into specified categories or clusters exhibiting similar characteristics in various groups. This can be further used to draw important interpretations in EDA.

Exploratory Data Analysis Tools

1. Python

Python is used for different tasks in EDA, such as finding missing values in data collection, data description, handling outliers, obtaining insights through charts, etc. The syntax for EDA libraries like Matplotlib, Pandas, Seaborn, NumPy, Altair, and more in Python is fairly simple and easy to use for beginners. You can find many open-source packages in Python, such as D-Tale, AutoViz, PandasProfiling, etc., that can automate the entire exploratory data analysis process and save time.

2. R

R programming language is a regularly used option to make statistical observations and analyze data, i.e., perform detailed EDA by data scientists and statisticians. Like Python, R is also an open-source programming language suitable for statistical computing and graphics. Apart from the commonly used libraries like ggplot, Leaflet, and Lattice, there are several powerful R libraries for automated EDA, such as Data Explorer, SmartEDA, GGally, etc.

3. MATLAB

MATLAB is a well-known commercial tool among engineers since it has a very strong mathematical calculation ability. Due to this, it is possible to use MATLAB for EDA but it requires some basic knowledge of the MATLAB programming language.

Advantages of Using EDA

Here are a few advantages of using Exploratory Data Analysis -

1. Gain Insights Into Underlying Trends and Patterns

EDA assists data analysts in identifying crucial trends quickly through data visualizations using various graphs, such as box plots and histograms. Businesses also expect to make some unexpected discoveries in the data while performing EDA, which can help improve certain existing business strategies.

2. Improved Understanding of Variables

Data analysts can significantly improve their comprehension of many factors related to the dataset. Using EDA, they can extract various information such as averages, means, minimum and maximum, and more such information is required for preprocessing the data appropriately.

3. Better Preprocess Data to Save Time

EDA can assist data analysts in identifying significant mistakes, abnormalities, or missing values in the existing dataset. Handling the above entities is critical for any organization before beginning a full study as it ensures correct preprocessing of data and may help save a significant amount of time by avoiding mistakes later when applying machine learning models.

4. Make Data-driven Decisions

The most significant advantage of employing EDA in an organization is that it helps businesses to improve their understanding of data. With EDA, they can use the available tools to extract critical insights and make conclusions, which assist in making decisions based on the insights from the EDA.

Example of Exploratory Data Analysis

Example 1: EDA in Health Care Research

Example 2: EDA in Retail

In the retail industry, EDA can be performed on a dataset consisting of various columns such as product categories, sales, price, discounts, region of sales, orders, etc., for understanding sales patterns, improving inventory management, predicting future demands, etc.

Example 3: EDA in Electronic Medical Records

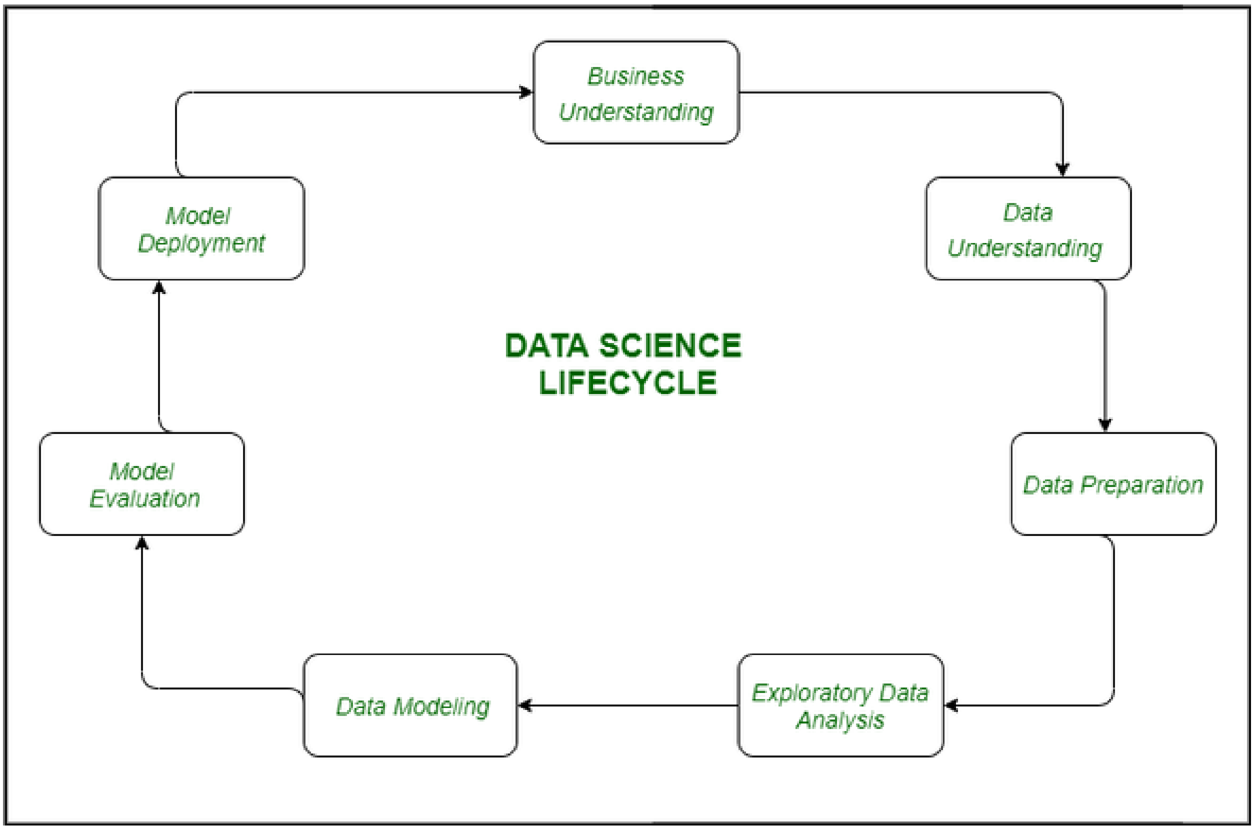
An important aspect for organizations in the healthcare domain is maintaining electronic medical records. These are digital records of the medical history of the visiting patients, such as any previous hospitalization, administered medicines, allergies or vaccinations,

The Data Science Process

Data Science Process Life Cycle

There are some steps that are necessary for any of the tasks which are being done in the field of data science to derive any fruitful results from the data at hand.

- **Data Collection** – After formulating any problem statement the main task is to calculate data that can help us in our analysis and manipulation. Sometimes data is collected by performing some kind of survey and there are times when it is done by performing scrapping.
- **Data Cleaning** – Most of the real-world data is not structured and requires cleaning and conversion into structured data before it can be used for any analysis or modeling.
- **Exploratory Data Analysis** – This is the step in which we try to find the hidden patterns in the data at hand. Also, we try to analyze different factors which affect the target variable and the extent to which it does so. How the independent features are related to each other and what can be done to achieve the desired results all these answers can be extracted from this process as well. This also gives us a direction in which we should work to get started with the modeling process.
- **Model Building** – Different types of machine learning algorithms as well as techniques have been developed which can easily identify complex patterns in the data which will be a very tedious task to be done by a human.
- **Model Deployment** – After a model is developed and gives better results on the holdout or the real-world dataset then we deploy it and monitor its performance. This is the main part where we use our learning from the data to be applied in real-world applications and use cases.



Components of Data Science Process

The main components of Data Science :

- **Data Analysis** –, we first perform an exploratory data analysis to get a basic idea of the data and patterns which are available in it this gives us a direction to work on if we want to apply some complex analysis methods on our data.
- **Statistics** – It is a natural phenomenon that many real-life datasets follow a normal distribution. And when we already know that a particular dataset follows some known distribution then most of its properties can be analyzed at once. Also, descriptive statistics and correlation and covariances between two features of the dataset help us get a better understanding of how one factor is related to the other in our dataset.
- **Data Engineering** – When we deal with a large amount of data then we have to make sure that the data is kept safe from any online threats also it is easy to retrieve and make changes in the data as well. To ensure that the data is used efficiently Data Engineers play a crucial role.

6 key steps of the data science life cycle explained

- Problem identification.
- Data investigation.
- Pre-processing of data.
- Exploratory data analysis.
- Data modeling.
- Model evaluation/ Monitoring.

01-Oct-2022

Follow these steps to accomplish your data science life cycle

1. Problem identification

Before you start your data science project, you need to identify the problem and its effects on patients. You can do this by conducting research on various sources, including:
minimize savings loss or prefers to predict the rate of a commodity.
To be precise, in this step we answer the following questions:

- Clearly state the problem to be solved
- Reason to solve the problem
- State the potential value of the project to motivate everyone
- Identify the stakeholders and risks associated with the project
- Perform high-level research with your data science team

Determine and communicate the project plan

2. Data investigation

In this step, we:

- Describe the data
- Define its structure
- Figure out relevance of data and
- Assess the type of data record

3. Pre-processing of data

The actions to be performed at this stage of a data science project are:

- Selection of the applicable data
- Data integration by means of merging the data sets
- Data cleaning and filtration of relevant information
- Treating the lacking values through either eliminating them or imputing them
- Treating inaccurate data through eliminating them
- Additionally, test for outliers the use of box plots and cope with them

This step also emphasizes the importance of elements essential to constructing new data

4. Exploratory data analysis

The following steps to conduct the Exploratory Data Analysis:

- Examine the data by formulating the various statistical functions
- Identify dependent and independent variables or features
- Analyze key features of data to work on
- Define the spread of data

5. Data modeling

Data modeling refers to the process of converting raw data into a form that can be transverse into other applications as well. Mostly, this step is performed in spreadsheets, but data scientists also prefer to use statistical tools and databases for data modeling.

The following elements are required for data modeling:

Data dictionary: A list of all the properties describing your data that you want to maintain in your system, for example, spreadsheet, database, or statistical software.

Entity relationship diagram: This diagram shows the relationship between entities in your data model. It shows how each element is related to the other, as well as any constraints to that relationship

Data model: A set of classes representing each piece of information in your system, along with its attributes and relationships with other objects in the system.

6. Model evaluation/ Monitoring

we need to know that model evaluation can be done parallel to the other stages of the data science life cycle. It helps you to know at every step if your model is working as intended or if you need to make any changes. Alongside, eradicate any error at an early stage to avoid getting false predictions at the end of the project.

In case you fail to acquire a quality result in the evaluation, we must reiterate the complete modeling procedure until the preferred stage of metrics is achieved.

Casestudy

2. Data Science in Healthcare

The Healthcare sector is immensely benefiting from the advancements in AI. Data science, especially in medical imaging, has been helping healthcare professionals come up with better diagnoses and effective treatments for patients. Similarly, several advanced healthcare analytics tools have been developed to generate clinical insights for improving patient care. These tools also assist in defining personalized medications for patients reducing operating costs for clinics and hospitals. Apart from medical imaging or computer vision, Natural Language Processing (NLP) is frequently used in the healthcare domain to study the published textual research data.

Pharmaceutical

Driving innovation with NLP: Novo Nordisk

Novo Nordisk uses the Linguamatics NLP platform from internal and external data sources for text mining purposes that include scientific abstracts, patents, grants, news, tech transfer offices from universities worldwide, and more. These NLP queries run across sources for the key therapeutic areas of interest to the Novo Nordisk R&D community. Several NLP algorithms have been developed for the topics of safety, efficacy, randomized controlled trials, patient populations, dosing, and devices. Novo Nordisk employs a data pipeline to capitalize the tools' success on real-world data and uses interactive dashboards and cloud services to visualize this standardized structured information from the queries for exploring commercial effectiveness, market situations, potential, and gaps in the product documentation. Through data science, they are able to automate the process of generating insights, save time and provide better insights for evidence-based decision making.

BioTech

How AstraZeneca harnesses data for innovation in medicine

AstraZeneca is a globally known biotech company that leverages data using AI technology to discover and deliver newer effective medicines faster. Within their R&D teams, they are using AI to decode the big data to understand better diseases like cancer, respiratory disease, and heart, kidney, and metabolic diseases to be effectively treated. Using data science, they can identify new targets for innovative medications. In 2021, they selected the first two AI-generated drug targets collaborating with BenevolentAI in Chronic Kidney Disease and Idiopathic Pulmonary Fibrosis.

Data science is also helping AstraZeneca redesign better clinical trials, achieve personalized medication strategies, and innovate the process of developing new medicines. Their Center for Genomics Research uses data science and AI to analyze around two million genomes by 2026. Apart from this, they are training their AI systems to check these images for disease and biomarkers for effective medicines for imaging purposes. This approach helps them analyze samples accurately and more effortlessly. Moreover, it can cut the analysis time by around 30%.

AstraZeneca also utilizes AI and machine learning to optimize the process at different stages and minimize the overall time for the clinical trials by analyzing the clinical trial data. Summing up, they use data science to design smarter clinical trials, develop innovative medicines, improve drug development and patient care strategies, and many more.

Wearable Technology

Wearable technology is a multi-billion-dollar industry. With an increasing awareness about fitness and nutrition, more individuals now prefer using fitness wearables to track their routines and lifestyle choices.

Fitness wearables are convenient to use, assist users in tracking their health, and encourage them to lead a healthier lifestyle. The medical devices in this domain are beneficial since they help monitor the patient's condition and communicate in an emergency situation. The regularly used fitness trackers and smartwatches from renowned companies like Garmin, Apple, FitBit, etc., continuously collect physiological data of the individuals wearing them. These wearable providers offer user-friendly dashboards to their customers for analyzing and tracking progress in their fitness journey.

3. Covid 19 and Data Science

In the past two years of the Pandemic, the power of data science has been more evident than ever. Different pharmaceutical companies across the globe could synthesize Covid 19 vaccines by analyzing the data to understand the trends and patterns of the outbreak. Data science made it possible to track the virus in real-time, predict patterns, devise effective strategies to fight the Pandemic, and many more.

How Johnson and Johnson uses data science to fight the Pandemic

The data science team at Johnson and Johnson leverages real-time data to track the spread of the virus. They built a global surveillance dashboard (granulated to county level) that helps them track the Pandemic's progress, predict potential hotspots of the virus, and narrow down the likely place where they should test its investigational COVID-19 vaccine candidate. The team works with in-country experts to determine whether official numbers are accurate and find the most valid information about case numbers, hospitalizations, mortality and testing rates, social compliance, and local policies to populate this dashboard. The team also studies the data to build models that help the company identify groups of individuals at risk of getting affected by the virus and explore effective treatments to improve patient outcomes.

4. Data Science in Ecommerce

In the e-commerce sector, big data analytics can assist in customer analysis, reduce operational costs, forecast trends for better sales, provide personalized shopping experiences to customers, and many more.

Amazon uses data science to personalize shopping experiences and improve customer satisfaction. Amazon is a globally leading eCommerce platform that offers a wide range of online shopping services. Due to this, Amazon generates a massive amount of data that can be leveraged to understand consumer behavior and generate insights on competitors' strategies. Amazon uses its data to provide recommendations to its users on different products and services. With this approach, Amazon is able to persuade its consumers into buying and making additional sales. This approach works well for Amazon as it earns 35% of the revenue yearly with this technique. Additionally, Amazon collects consumer data for faster order tracking and better deliveries.

Similarly, Amazon's virtual assistant, Alexa, can converse in different languages; uses speakers and a camera to interact with the users. Amazon utilizes the audio commands from users to improve Alexa and deliver a better user experience.

7. Data Science in Entertainment Industry

Due to the Pandemic, demand for OTT (Over-the-top) media platforms has grown significantly. People prefer watching movies and web series or listening to the music of their choice at leisure in the convenience of their homes. This sudden growth in demand has given rise to stiff competition. Every platform now uses data analytics in different capacities to provide better-personalized recommendations to its subscribers and improve user experience.

How Netflix uses data science to personalize the content and improve recommendations

Netflix is an extremely popular internet television platform with streamable content offered in several languages and caters to various audiences. In 2006, when Netflix entered this media streaming market, they were interested in increasing the efficiency of their existing "Cinematch" platform by 10% and hence, offered a prize of $1 million to the winning team. This approach was successful as they found a solution developed by the BellKor team at the end of the competition that increased prediction accuracy by 10.06%. Over 200 work hours and an ensemble of 107 algorithms provided this result. These winning algorithms are now a part of the Netflix recommendation system.

Netflix also employs Ranking Algorithms to generate personalized recommendations of movies and TV Shows appealing to its users.

Spotify uses big data to deliver a rich user experience for online music streaming

Personalized online music streaming is another area where data science is being used. Spotify is a well-known on-demand music service provider launched in 2008, which effectively leveraged big data to create personalized experiences for each user. It is a huge platform with more than 24 million subscribers and hosts a database of nearly 20million songs; they use the big data to offer a rich experience to its users. Spotify uses this big data and various algorithms to train machine learning models to provide personalized content. Spotify offers a "Discover Weekly" feature that generates a personalized playlist of fresh unheard songs matching the user's taste every week. Using the Spotify "Wrapped" feature, users get an overview of their most favorite or frequently listened songs during the entire year in December. Spotify also leverages the data to run targeted ads to grow its business. Thus, Spotify utilizes the user data, which is big data and some external data, to deliver a high-quality user experience

8. Data Science in Banking and Finance

Data science is extremely valuable in the Banking and Finance industry. Several high priority aspects of Banking and Finance like credit risk modeling (possibility of repayment of a loan), fraud detection (detection of malicious or irregularities in transactional patterns using machine learning), identifying customer lifetime value (prediction of bank performance based on existing and potential customers), customer segmentation (customer profiling based on behavior and characteristics for personalization of offers and services). Finally, data science is also used in real-time predictive analytics (computational techniques to predict future events).

How HDFC utilizes Big Data Analytics to increase revenues and enhance the banking experience

One of the major private banks in India, HDFC Bank, was an early adopter of AI. It started with Big Data analytics in 2004, intending to grow its revenue and understand its customers and markets better than its competitors. Back then, they were trendsetters by setting up an enterprise data warehouse in the bank to be able to track the differentiation to be given to customers based on their relationship value with HDFC Bank. Data science and analytics have been crucial in helping HDFC bank segregate its customers and offer customized personal or commercial banking services. The analytics engine and SaaS use have been assisting the HDFC bank in cross-selling relevant offers to its customers. Apart from the regular fraud prevention, it assists in keeping track of customer credit histories and has also been the reason for the speedy loan approvals offered by the bank.

Unit 4

Data Science and it's Various Applications

Data Science is the deep study of a large quantity of data, which involves extracting some meaningful from the raw, structured, and unstructured data. The extracting out meaningful data from large amounts use processing of data and this processing can be done using statistical techniques and algorithm, scientific techniques, different technologies, etc. It uses various tools and techniques to extract meaningful data from raw data. Data Science is also known as the **Future of Artificial Intelligence**.

For Example, Jagroop loves books to read but every time when he wants to buy some books he was always confused that which book he should buy as there are plenty of choices in front of him. This Data Science Technique will useful. When he opens Amazon he will get product recommendations on the basis of his previous data. When he chooses one of them he also gets a recommendation to buy these books with this one as this set is mostly bought. So all Recommendation of Products and Showing set of books purchased collectively is one of the examples of Data Science.

Applications of Data Science

1. In Search Engines

The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.

2. In Transport

Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.

For Example, In Driverless Cars the training data is fed into the algorithm and with the help of Data Science techniques, the Data is analyzed like what is the speed limit in Highway, Busy Streets, Narrow Roads, etc. And how to handle different situations while driving etc.

3. In Finance

Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company. Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.

For Example, In Stock Market, Data Science is the main part. In the Stock Market, Data Science is used to examine past behavior with past data and their goal is to examine the future outcome. Data is analyzed in such a way that it makes it possible to predict future stock prices over a set timetable.

4. In E-Commerce

E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.

For Example, When we search for something on the E-commerce websites we get suggestions similar to choices according to our past data and also we get recommendations according to most buy the product, most rated, most searched, etc. This is all done with the help of Data Science.

5. In Health Care

In the Healthcare Industry data science act as a boon. Data Science is used for:

- Detecting Tumor.
- Drug discoveries.
- Medical Image Analysis.
- Virtual Medical Bots.
- Genetics and Genomics.
- Predictive Modeling for Diagnosis etc.

6. Image Recognition

Currently, Data Science is also used in Image Recognition. **For Example,** When we upload our image with our friend on Facebook, Facebook gives suggestions Tagging who is in the picture. This is done with the help of machine learning and Data Science. When an Image is Recognized, the data analysis is done on one’s Facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile then Facebook suggests us auto-tagging.

7. Targeting Recommendation

Targeting Recommendation is the most important application of Data Science. Whatever the user searches on the Internet, he/she will see numerous posts everywhere. This can be explained properly with an example: Suppose I want a mobile phone, so I just Google search it and after that, I changed my mind to buy offline. Data Science helps those companies who are paying for Advertisements for their mobile. So everywhere on the internet in the social media, in the websites, in the apps everywhere I will see the recommendation of that mobile phone which I searched for. So this will force me to buy online.

8. Airline Routing Planning

With the help of Data Science, Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or take a halt in between like a flight can have a direct route from Delhi to the U.S.A or it can halt in between after that reach at the destination.

9. Data Science in Gaming

In most of the games where a user will play with an opponent i.e. a Computer Opponent, data science concepts are used with machine learning where with the help of past data the Computer will improve its performance. There are many games like Chess, EA Sports, etc. will use Data Science concepts.

10. Medicine and Drug Development

The process of creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone's life. Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors. The algorithms based on data science will forecast how this will react to the human body without lab experiments.

11. In Delivery Logistics

Various Logistics companies like DHL, FedEx, etc. make use of Data Science. Data Science helps these companies to find the best route for the Shipment of their Products, the best time suited for delivery, the best mode of transport to reach the destination, etc.

12. Autocomplete

AutoComplete feature is an important part of Data Science where the user will get the facility to just type a few letters or words, and he will get the feature of auto-completing the line. In Google Mail, when we are writing formal mail to someone so at that time data science concept of Autocomplete feature is used where he/she is an efficient choice to auto-complete the whole line. Also in Search Engines in social media, in various apps, AutoComplete feature is widely used.

ta scientists tackle questions about the future. They start with big data, characterized by the three V’s: volume, variety and velocity. Then, they use it as fodder for algorithms and models. The most cutting-edge data scientists, working in machine learning and AI, make models that automatically self-improve, noting and learning from their mistakes.

Data scientists have changed almost every industry. In medicine, their algorithms help predict patient side effects. In sports, their models and metrics have redefined “athletic potential.” Data science applications have even tackled traffic, with route-optimization models that capture typical rush hours and weekend lulls.

Below we’ve rounded up 25 examples of data science applications at work, in areas from e-commerce to healthcare.

DATA SCIENCE APPLICATIONS AND EXAMPLES

- **Healthcare:** Data science can identify and predict disease, and personalize healthcare recommendations.
- **Transportation:** Data science can optimize shipping routes in real-time.
- **Sports:** Data science can accurately evaluate athletes’ performance.
- **Government:** Data science can prevent tax evasion and predict incarceration rates.
- **E-commerce:** Data science can automate digital ad placement.
- **Gaming:** Data science can improve online gaming experiences.
- **Social media:** Data science can create algorithms to pinpoint compatible partners.
- **Fintech:** Data science can help create credit reports and financial profiles, run accelerated underwriting and create predictive models based on historical payroll data.

Healthcare Data Science Applications

Back in 2008, data science made its first major mark on the healthcare industry. Google staffers discovered they could map flu outbreaks in real time by tracking location data on flu-related searches. The CDC's existing maps of documented flu cases, FluView, was updated only once a week. Google quickly rolled out a competing tool with more frequent updates: Google Flu Trends.

But it didn't work. In 2013, Google estimated about twice the flu cases that were actually observed. The tool's secret methodology seemed to involve finding correlations between search term volume and flu cases. That meant the Flu Trends algorithm sometimes put too much stock in seasonal search terms like "high school basketball."

Even so, it demonstrated the serious potential of data science in healthcare. Here are some examples of more powerful and precise healthcare tools developed in the years after Google's initial attempt. All of them are powered by data science.

1. IDENTIFYING CANCER TUMORS

Google hasn't abandoned applying data science to healthcare. In fact, the company developed a tool, LYNA, for identifying breast cancer tumors that metastasize to nearby lymph nodes. That can be difficult for the human eye to see, especially when the new cancer growth is small. In one trial, LYNA — short for Lymph Node Assistant —accurately identified metastatic cancer 99 percent of the time using its machine-learning algorithm. More testing is required, however, before doctors can use it in hospitals.

2. TRACKING MENSTRUAL CYCLES

The popular Clue app employs data science to forecast users' menstrual cycles and reproductive health by tracking cycle start dates, moods, stool type, hair condition and many other metrics. Behind the scenes, data scientists mine this wealth of anonymized data with tools like Python and Jupyter's Notebook. Users are then algorithmically notified when they're fertile, on the cusp of a period or at an elevated risk for conditions like an ectopic pregnancy.

3. PERSONALIZING TREATMENT PLANS

Oncora's software uses machine learning to create personalized recommendations for current cancer patients based on data from past ones. Healthcare facilities using the company's platform include UT Health San Antonio and Scripps Health. Their radiology team collaborated with Oncora data scientists to mine 15 years' worth of data on diagnoses, treatment plans, outcomes and side effects from more than 50,000 cancer records. Based on this data, Oncora's algorithm learned to suggest personalized chemotherapy and radiation regimens.

4. CLEANING CLINICAL TRIAL DATA

Veeva is a cloud software company that provides data and software solutions for the healthcare industry. The company's reach extends through clinical, regulatory and commercial medical fields. Veeva's Vault EDC uses data science to clean clinical trial findings and help medical professionals make adjustments mid-study.

RELATED READINGData Science Versus Computer Science: What's the Difference?

Transportation and Logistics Data Science Examples

Driving plays a central role in American life. The Supreme Court has called it "a virtual necessity," and the vast majority of Americans — about 132 million households — own or lease cars. In 2021, American automobiles burned about 135 billion gallons of gasoline. Unfortunately, this habit contributes to climate change, but data science is here to help.

While both biking and public transit can curb driving-related emissions, data science can do the same by optimizing road routes. And though data-driven route adjustments are often small, they can help save thousands of gallons of gas when spread across hundreds of trips and vehicles — even among companies that aren't explicitly eco-focused. Here are some examples of data science hitting the road.

5. MODELING TRAFFIC PATTERNS

StreetLight uses data science to model traffic patterns for cars, bikes and pedestrians on North American streets. Based on a monthly influx of trillions of data points from smartphones, in-vehicle navigation devices and more, Streetlight's traffic maps stay up-to-date. They're more granular than mainstream maps apps too: they can identify groups of commuters that use multiple transit modes to get to work, like a train followed by a scooter. The company's maps inform various city planning enterprises, including commuter transit design.

6. OPTIMIZING FOOD DELIVERY

The data scientists at UberEats have a fairly simple goal: getting hot food delivered quickly. Making that happen across the country though, takes machine learning, advanced statistical modeling and staff meteorologists. In order to optimize the full delivery process, the team has to predict how every possible variable — from storms to holiday rushes — will impact traffic and cooking time.

7. IMPROVING PACKAGE DELIVERY

UPS uses data science to optimize package transport from drop-off to delivery. The company's integrated navigation system ORION helps drivers choose over 66,000 fuel-efficient routes. ORION has saved UPS approximately 100 million miles and 10 million gallons of fuel per year with the use of advanced algorithms, AI and machine learning. The company plans to continue to update its ORION system, with the last version having been rolled out in 2021. The latest update allowed drivers to reduce their routes by two to four miles.

Sports Data Science Applications

In the early 2000s, the Oakland Athletics' recruitment budget was so small the team couldn't recruit quality players. At least, they couldn't recruit players any other teams considered quality. So the general manager redefined quality, using in-game statistics other teams ignored to predict player potential and assemble a strong team despite their budget.

His strategy helped the A's make the playoffs, and it snowballed from there. Author Michael Lewis wrote a book about the phenomenon, *Moneyball*. Since then, the global market for sports analytics has grown significantly and is expected to reach 8.4 billion by 2026. Here are some examples of how data science is transforming sports.

8. MAKING PREDICTIVE INSIGHTS IN BASKETBALL

RSPCT's shooting analysis system, adopted by NBA and college teams, relies on a sensor on a basketball hoop's rim, whose tiny camera tracks exactly when and where the ball strikes on each basket attempt. It funnels that data to a device that displays shot details in real time and generates predictive insights.

“Based on our data… We can tell [a shooter], ‘If you are about to take the last shot to win the game, don’t take it from the top of the key, because your best location is actually the right corner,’” RSPCT COO Leo Moravtchik told SVG News.

9. TRACKING PHYSICAL DATA FOR ATHLETES

WHOOP makes wearable devices that track athletes’ physical data like resting heart rate, sleep cycle and respiratory rate. The goal is to help athletes understand when to push their training and when to rest — and to make sure they’re taking the necessary steps to get the most out of their body. Professional athletes like Olympic sprinter Gabby Thomas, Olympic golfer Nelly Korda and PGA golfer Nick Watney are among the WHOOPS’ users, according to the company’s website.

10. GATHERING PERFORMANCE METRICS FOR SOCCER PLAYERS

Trace provides soccer coaches with recording gear and an AI system that analyzes game film. Players wear a tracking device, called a Tracer, while its specially designed camera records the game. The AI bot then takes that footage and stitches together all of the most important moments in a game — from shots on goal to defensive lapses and more. This technology allows coaches and players to have more detailed insights from game film. Beyond stitching together clips, the software also provides performance metrics and a field heat map.

MORE ON DATA SCIENCE IN SPORTSHow Data Science and Analytics Came to Dominate Fantasy Football

Government Data Science Applications

Though few think of the U.S. government as “extremely online,” its agencies can access heaps of data. Not only do its agencies maintain their own databases of ID photos, fingerprints and phone activity, government agents can get warrants to obtain data from any American data warehouse. Investigators often reach out to Google's warehouse, for instance, to get a list of the devices that were active at the scene of a crime.

Though many view such activity as an invasion of privacy, the United States has minimal privacy regulations, and the government's data well won't run dry anytime soon. Here are some of the ways government agencies apply data science to vast stores of data.

12. MINING DATABASES WITH FACIAL RECOGNITION SOFTWARE

The U.S. Immigrations and Customs Enforcement has used facial recognition technology to mine driver's license photo databases, with the goal of deporting undocumented immigrants. The practice — which has sparked criticism from both an ethical and technological standpoint (facial recognition technology remains shaky) — falls under the umbrella of data science. Facial recognition builds on photos of faces, a.k.a raw data, with AI and machine learning capabilities.

13. DETECTING TAX FRAUD

Tax evasion costs the U.S. government $1 tillion a year, according to one estimate, so it's no wonder the IRS has modernized its fraud-detection protocols in the digital age. To the dismay of privacy advocates, the agency has improved efficiency by constructing multidimensional taxpayer profiles from public social media data, assorted metadata, emailing analysis, electronic payment patterns and more. Based on those profiles, the agency forecasts individual tax returns; anyone with wildly different real and forecasted returns gets flagged for auditing.

Gaming Data Science Examples

The gaming industry is growing, and it’s using data science to help expand. The global video game market was valued at $195.65 billion in 2021 and is expected to grow by nearly 13 percent by 2030.

Data science and AI have been used in video games since as early as the 1950s with the creation of Nim — a mathematical strategy game in which two players take turns to remove objects from piles. The innovation continued with Pac-Man where AI and data science were used in the game’s mazes and to give the ghosts distinct personalities.

The video game industry continues to find creative ways to implement data science and AI to improve game play and entertain millions of people across the globe. Here are just a few examples of how data science is used in video games.

14. IMPROVING ONLINE GAMING

Known for being the company behind games with cult followings like Call of Duty, World of Warcraft, Candy Crush and Overwatch, Activision Blizzard uses big data to improve their online gaming experiences. One example of this being the company’s game science division analyzing gaming data to prevent empowerment — the attempt to improve someone else’s sports scores through negative means — amongst COD players. The company also uses machine learning to detect power boosting and identify and track key indicators for increasing quality of game time.

15. MAKING SUGGESTIONS TO GAMERS TO IMPROVE PLAY

2k Games is a video game studio that has created popular titles like Bioshock and Borderlands, as well as both WWE and PGA games series. The company's growing game science team focuses on extracting gaming data and building models in order to improve its sports games like NBA2K. Data scientists at 2K games analyze player gameplay and economy telemetry data to understand player behavior and suggest actions to improve the player experience.

16. MONITORING BUSINESS METRICS IN THE VIDEO GAME INDUSTRY

Unity is a platform for creating and operating interactive, real-time 3D content, including games. The platform is used by gaming companies like Riot Games, Atari and Respawn Entertainment, according to its website. Unity uses gaming data to make data-driven decision making within its product development team and to monitor business metrics.

E-Commerce Data Science Applications

Once upon a time, everyone in a given town shopped at the same mall: a physical place with some indoor fountains, a jewelry kiosk and probably a Body Shop. Today, citizens of that same town can each shop in their own personalized digital mall — also known as the internet. Online retailers often automatically tailor their web storefronts based on viewers' data profiles. That can mean tweaking page layouts and customizing spotlighted products, among other things. Some stores may also adjust prices based on what consumers seem able to pay, a practice called personalized pricing. Even websites that sell nothing feature targeted ads. Here are some examples of companies using data science to automatically personalize the online shopping experience.

17. CREATING TARGETED ADS

Sovrn brokers deals between advertisers and outlets like Bustle, ESPN and Encyclopedia Britannica. Since these deals happen millions of times a day, Sovrn has mined a lot of data for insights, which manifest in its intelligent advertising technology. Compatible with Google and Amazon's server-to-server bidding platforms, its interface can monetize media with minimal human oversight — or, on the advertiser end, target campaigns to customers with specific intentions.

18. CURATING VACATION RENTALS

Data science helped Airbnb totally revamp its search function. Once upon a time, it prioritized top-rated vacation rentals that were located a certain distance from a city's center. That meant users could always find beautiful rentals, but not always in cool neighborhoods. Engineers solved that issue by prioritizing the search rankings of a rental if it's in an area that has a high density of Airbnb bookings. There's still breathing room for quirkiness in the algorithm, too, so cities don't dominate towns and users can stumble on the occasional rental treehouse.

19. PREDICTING CONSUMERS' PRODUCT INTERESTS

Instagram uses data science to target its sponsored posts, which hawk everything from trendy sneakers to influencers posting sponsored ads. The company's data scientists pull data from Instagram as well as its owner, Meta, which has exhaustive web-tracking infrastructure and detailed information on many users, including age and education. From there, the team crafts algorithms that convert users' likes and comments, their usage of other apps and their web history into predictions about the products they might buy.

Though Instagram's advertising algorithms remain shrouded in mystery, they work impressively well, according to *The Atlantic*'s Amanda Mull: "I often feel like Instagram isn't pushing products, but acting as a digital personal shopper I'm free to command."

20. CREATING DIGITAL AD OPPORTUNITIES

Taboola uses deep learning, AI and large datasets to create engagement opportunities for advertisers and digital properties. Its discovery platform creates new monetization, audience and engagement by placing advertisements throughout a variety of online publishers and sites. Its discovery platform can expose readers to news, entertainment, topical information or advice as well as a new product or service. The company partners with outlets like *USA Today*, *Bloomberg*, *Insider* and MSN, according to its website.

Social Platform Data Science Examples

The rise of social networks has completely altered how people socialize. Romantic relationships unfold publicly on Venmo. Meta engineers can rifle through users' birthday party invite lists. Friendship, acquaintanceship and coworker-ship all leave extensive online data trails.

Some argue that these trails — Facebook friend lists or LinkedIn connections — don't mean much. Anthropologist Robin Dunbar, for instance, has found that people can maintain only about 150 casual connections at a time; cognitively, humans can't handle much more than that. In Dunbar's view, racking up more than 150 digital connections says little about a person's day-to-day social life.

Catalogs of social network users' most glancing acquaintances hold another kind of significance though. Now that many relationships begin online, data about your social world impacts who you get to know next. Here are some examples of data science fostering human connection.

(UNIT-5)

Data Science Tool Kit

Data science toolkit is a list of functions, modules, packages, frameworks, software that can really help a data scientist to solve a problem. Sometimes these functions and packages available in form of 3rd party packages or software and sometimes you are required to create your own. That's why a True Data Scientist is a mix of Statistician and a Programmer .

SaS : Statistical Analytical System.

Statistical Analysis System (SAS) is an integrated system of software products provided by SAS Institute Inc., which enables programmers to perform:

- Information retrieval and data management
- Report writing and graphics
- Statistical analysis, econometrics and data mining
- Business planning, forecasting, and decision support
- Operations research and project management
- Quality improvement
- Applications development
- Data warehousing (extract, transform, load)
- Platform independent and remote computing

It is a tool developed for advanced analytics and complex statistical operations. It is used by large scale organizations and professionals due to its high reliability.

Apache Spark :

Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters. Simple. Fast.

Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance. Originally developed at the University of California, Berkeley's AMPLab, the Spark codebase was later donated to the Apache Software Foundation, which has maintained it since.

BigML :

BigML is a consumable, programmable, and scalable Machine Learning platform that makes it easy to solve and automate Classification, Regression, Time Series Forecasting, Cluster Analysis, Anomaly Detection, Association Discovery, and Topic Modeling tasks.

Excel **:** Excel is used in developing simple level applications recommendations, fraud detections etc., to the complex level applications like building Self Driving Cars. It is recognized as the most powerful tool of Data Science.

R-Programming :

R is an open-source programming language that is widely used as a statistical software and data analysis tool. R is an important tool for Data Science. It is highly popular and is the first choice of many statisticians and data scientists.

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac.

R is free software distributed under a GNU-style copy left, and an official part of the GNU project called **GNU S**.

TensorFlow :

The Tensorflow framework is an open end-to-end machine learning platform. It's a symbolic math toolkit that integrates data flow and differentiable programming to handle various tasks related to deep neural network training and inference.

KNIME : Konstanz Information Miner,

is a free and open-source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining "Building Blocks of Analytics" concept.

Tableau:

Tableau is a collection of various Business Intelligence and data analytics tools that allows the user to collect data from varied sources in both structured and unstructured format and convert that data into visualizations and other insights.

PowerBI :

Power BI is a technology-driven business intelligence tool provided by Microsoft for analyzing and visualizing raw data to present actionable information. It combines business analytics, data visualization, and best practices that help an organization to make data-driven decisions.

Power BI supports both M Language and DAX as expression languages. Both are more comparable to the formulas in Microsoft Excel than they are to any programming language. However, M and DAX are distinct from one another and are applied in various ways when creating Power BI models