Unit-1

Introduction:

Data science is a multidisciplinary field that combines various techniques, algorithms, processes, and systems to extract valuable insights and knowledge from data.

It involves collecting, analyzing, interpreting, and visualizing data to make informed decisions and predictions.

Data science typically incorporates elements of statistics, computer science, machine learning, and domain expertise to tackle complex problems and discover patterns, trends, and meaningful information in large datasets.

It has applications in various industries, such as business, healthcare, finance, and more, to help organizations optimize operations, improve decision-making, and drive innovation.

In essence, data science is all about transforming raw data into actionable intelligence.

Big Data:

Big Data refers to extremely large and complex datasets that are beyond the capacity of traditional data processing methods to handle effectively. These datasets are characterized by their volume, velocity, variety, and often, veracity.

-> Volume: Big Data involves vast amounts of information, sometimes ranging from terabytes to exabytes, which is much larger than what can be managed using conventional databases and tools.

-> Velocity: Data is generated and updated at a high speed in real-time or near real-time, such as social media posts, sensor data, or financial transactions.

-> Variety: Big Data comprises a diverse range of data types, including structured (e.g., databases), semi-structured (e.g., XML), and unstructured (e.g., text, images, videos).

-> Veracity: Big Data can be noisy and of varying quality, which means that not all the data within a dataset is accurate or reliable.

Hype of Data Science:

The hype surrounding data science is driven by the promise of valuable insights from Big Data, technological advancements, a growing job market, the influence of successful data-driven companies, and media coverage.

The hype surrounding "data science" can be attributed to several factors:

1. Potential for Insights: Data science has the potential to extract valuable insights and patterns from massive datasets. Businesses and organizations see this as an opportunity to gain a competitive advantage and make more informed decisions.

2. Technological Advancements: Advancements in computing power, storage, and software tools have made it possible to process and analyze Big Data more efficiently.

3. Market Demand: The job market for data scientists has grown significantly, and high demand for professionals in this field has led to increased attention and investment.

4. Success Stories: Success stories of data-driven companies like Google, Facebook, and Amazon have drawn considerable attention to data science's potential for innovation and revenue generation.

5. Media Coverage: The media often portrays data scientists as "Masters of the Universe," creating a mystique around the field that contributes to the hype.

6. Applications Across Industries: Data science is applicable in various industries, including healthcare, finance, marketing, and more, making it a versatile and widely discussed topic.

7. Ambiguity of Terminology: As the introduction mentioned, there's a lack of precise definitions for terms like "data science" and "Big Data," which can lead to misconceptions and further fuel the hype.

Datafication:

Datafication is the process of transforming various aspects of life, actions, behaviors, and interactions into digital data.

It involves the collection, recording, and analysis of activities and information to create structured digital records.

Datafication often occurs through the use of digital devices, sensors, and software that capture and convert real-world events and behaviors into data, which can then be stored, processed, and analyzed for various purposes.

This process allows for the quantification and measurement of a wide range of activities, from social interactions on social media to tracking physical movements through location data, with the goal of deriving insights, making predictions, and creating value from the collected data.

Datafication raises the issue of intentionality in sharing data. When people "like" something online, they are willingly participating in datafication. However, merely browsing the web can unintentionally lead to datafication through cookies and tracking.

Datafication can also occur without a person's knowledge, such as when sensors, cameras, or devices like Google Glasses collect data as people move around in physical spaces.

The ability to datafy things enables the transformation of information into new forms of value. This value is often associated with increased efficiency and automation, with businesses and entrepreneurs making money by leveraging data.

Statistical Inference:

Statistics is a collection of quantitative (measurable) and numeral data which is suitable for analysis.

Ex: height, no.of emails sent by employees per year, etc.

Inference is the process of making a decision/ drawing a conclusion, based on the provided evidence.

Hence, the process of making decisions/ drawing a conclusion, based on the provided evidence by the analysis of quantitative and numeral data sets, is known as statistical inference.

This is of two types:

(1) Descriptive Statistics

(2) Inferential Statistics

(1) Descriptive Statistics: The statistics which summarize the data and describe the main features of dataset. The includes measures such as mean, median, mode, variance, and standard deviation.

Ex: Average mark of a class (mean), frequently repeated mark (mode), center mark (mode).

(2) Inferential Statistics: The statistics which draw inferences (conclusions) from the summarized data. This includes hypothesis testing, confidence intervals, and regression analysis.

Ex: Are the marks of students of a class similar to other classes.

Population and Samples:

An entire set of observations/elements/objects/units that share common characteristics and is used as the data to make statistical inferences, is known as Population.

A subset of a population is known as the sample.

A sample is used to draw conclusions about the whole population.

A sample must be similar to its population. (I.e., it must be drawn from a population whose conclusions are to be drawn.)

Ex:

Population	Sample
All of the people in Germany	500 Germans
All of the customers of Netflix	300 Netflix customers
Every car manufacturer	Tesla, Toyota, BMW, Ford

Statistical Modeling:

Modeling is the process of building a model (a copy/instance of something) by using finite data sets.

Statistical Modeling is the process of building a model, by using statistical methods on the finite data sets provided.

These models can be used for:

- Create a representation of data.
- Perform analysis to deduce correlations between variables. (for clustering and labeling the similar data points)
- Uncover insights.
- Generate sample data.
- Make predictions about the real world.

Probability Distributions:

Fitting a Model:

A model is a computer program built with complex algorithms and a bunch of training data that can be used to recognize patterns in provided data and make predictions.

That is, initially training data is provided to the algorithm to improve and optimize.

After which, real-world data (test data) is provided to extract useful patterns and make predictions from it.

These patterns and predictions can be used in various fields, such as business intelligence, whether forecast, etc.

Model Fitting is a measure of how well a model generalizes(understands) the test data which is similar to the training data provided to it.

That is, the process of measuring the efficiency and accuracy of a model on a particular type of data set.

There are three types of fittings:

(1) Overfitting

(2) Underfitting

(3) Bestfitting

(1) Overfitting:

-> When a model performs very well for training data but has poor performance with test data (new data), it is known as overfitting.

-> In this case, the machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data.

-> Overfitting can happen due to low bias and high variance.

(2) Underfitting:

(3)Bestfitting:

Basics Of R:

Introduction:

R is a popular programming language used for statistical computing and graphical presentation.

Its most common use is to analyze and visualize data.

There are other programming languages for the same purposes, such as Python and MATLAB.

So, the major reasons for using R, are:

- It is a great resource for data analysis, data visualization, data science and machine learning
- It provides many statistical techniques (such as statistical tests, classification, clustering and data reduction)
- It is easy to draw graphs in R, like pie charts, histograms, box plot, scatter plot, etc++
- It works on different platforms (Windows, Mac, Linux)
- It is open-source and free
- It has a large community support
- It has many packages (libraries of functions) that can be used to solve different problems

Features of R:

1. Open-Source

You don't have to pay any money to download R on your computer. It is free and open-source software. Furthermore, you can contribute towards the development of R, customize its packages, and add more features.

2. Strong Ability to Design Graphics

R has improved libraries that make it possible to create interactive graphics. As a result, data visualization and representation are relatively simple. R can generate various flow diagrams, from straightforward charts to intricate, interactive ones.

3. Extensive Range of Packages

CRAN, or the Comprehensive R Archive Network, contains over **10,000** different packages and extensions that help handle a wide range of data science challenges. R contains a large set of packages for many subjects, such as astronomy, biology, and so forth. While R was developed for academic objectives, it is now also utilized in industry.

4. Efficient in Software Development

R has an extensive development environment, which means it may be used for both statistical computing and software development. R is an **object-oriented programming** language. It also includes a powerful package called Rshiny that can be used to create full-fledged web apps.

5. Computing in a Distributed Environment

Tasks are split across numerous processing nodes in distributed computing to minimize processing time and boost efficiency. R offers tools like **"ddR"** and **"multiDplyr"** that allow it to process big data sets using distributed computing.

6. Data Wrangling

The process of cleansing large and inconsistent data sets in order to facilitate computation and further analysis is known as **data wrangling**. This is a time-consuming process. R's broad tool collection can be utilized for database management and wrangling.

7. No Compilation

The R language is interpreted rather than compiled. As a result, **no compiler is required** to compile code into an executable program. The R code is evaluated step by step and turned straight into machine-level calls. This significantly reduces the time required to run a R script.

8. Enables Quick Calculations

R supports a wide range of complicated operations on **vectors, arrays, data frames,** and other data objects of various sizes. Furthermore, all of these actions occur at breakneck speed. It includes a variety of operator suites to execute these varied calculations.

9. Integration with Other Technologies

Many other technologies, frameworks, software programs, and programming languages can be combined with R. To use Hadoop's distributed computing capabilities. It can be linked with it. Additionally, it may be integrated with programs written in **FORTRAN, C, C++, Java, and Python**, among other computer languages.

10. Compatibility with Multiple Platforms

R allows for **cross-platform compatibility**. It can run on any operating system and in any software environment. It can also run on any hardware setup without the need for any further workarounds.

Advantages of R:

- R is the most comprehensive statistical analysis package. As new technology and concepts often appear first in R.
- As R programming language is an open source. Thus, you can run R anywhere and at any time.
- R programming language is suitable for GNU/Linux and Windows operating system.
- R programming is a cross-platform which runs on any operating system.
- In R, everyone is welcome to provide new packages, bug fixes, and code enhancements.

Disadvantages of R:

- In the R programming language, the standard of some packages is less than perfect.
- Although, R commands put little pressure on memory management. So, R programming language may consume all available memory.
- In R basically, nobody complains if something doesn't work.
- R programming language is much slower than other programming languages such as Python and MATLAB.

R-Environmental Setup:

Programming with R:

Basic Data Types:
