**Unit 1**

**Chp. 1 - Introduction to Data Science**

**1.1 What is Data ?**

Data is a set of [characters](https://www.computerhope.com/jargon/c/charact.htm) that is gathered and translated for some purpose, usually Data analysis.

It can be numbers, any character, including text and, sound ,pictures or video.

Examples of computer data

Joe,Smith, 1234

0143 0157 01554

011000110110111

Data is defined as raw facts and figures collected together and stored in database.

Data can be in Structured ,semi-structured and unstructured format.

1. **Structured** - By structured data, we mean data that can be processed, stored, and

retrieved in a fixed format. It refers to highly organized information that can be readily and

seamlessly stored and accessed from a database by simple search engine algorithms**. For**

**instance, the employee table in a company database will be structured as the employee**

**details, their job positions, their salaries, etc.,** will be present in an organized manner.

#### Semi-structured- Semi-structured data pertains to the data containing both the

#### formats mentioned above, that is, structured and unstructured data. To be precise,

#### it refers to the data that although has not been classified under a particular

#### repository (database), yet contains vital information or tags that segregate individual

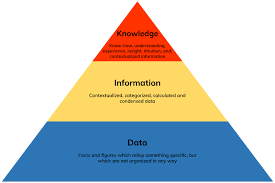
#### elements within the data.

#### Unstructured- Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data.

As we know data is a raw information, we need to process it, clean it to use for data analysis purpose.

Data can be measured , collected, presented and analyzed by using various analysis tools.

Following fig. Describes Data, Information and knowledge.

****

**Data :** Data is comprised of basic unrefined and generally unfiltered data information.

**Information :** Information is much more refined data that has evolved to the point of being useful for some form of analysis.

**Knowledge** : Knowledge resides in the user, happens only when human experiences and inside is applied to the data and information.

**1.2 Different kinds of data**

At the highest level, two kinds of data exists : ***Numerical*** and ***Categorial***.

1. **Numerical Data** : It is also referred as Quantitative Data. This include information that is measurable. Data represented as numbers and not words or text that we can measure in dimensions such as height, width, length, Temperature, humidity, Price , Area and Volume.
2. **Categorial Data** : It is also referred as Qualitative Data. This is any data that isn’t a number, which can mean a string of text or date. This data cannot be measured easily but can be observed such as smells, taste, colour etc.
3. ***Numerical Data***

Numerical Data is classified in two types : ***Discrete and Continuous***

***Discrete : Discrete*** data is a count that can't be made more precise. Typically it involves integers. For instance, the number of children (or adults, or pets) in your family is discrete data, because you are counting whole, indivisible entities: you can't have 2.5 kids, or 1.3 pets.

***Continuous : Continuous*** data, on the other hand, could be divided and reduced to finer and finer levels. For example, you can measure the height of your kids at progressively more precise scales—meters, centimeters, millimeters, and beyond—so height is continuous data.

1. **Categorial Data**

Categorial Data is classified in two types : **Ordinal,Nominal and Binary**

**Ordinal :** Ordinal data is ain which data have natural order. E.g height can have three categories short, medium and tall. Credit card have two categories low and high.

**Nominal** : Nominal data is used to label variables without providing any quantitative

value. It is simplest form of measure. Unlike ordinal data, nominal data cannot be

ordered and cannot be measured. Nominal value examples include variables such as

“Country” or “Marital Status”. Marital status can be single, Married or Divorced.

**Binary** : Binary data have only two values –yes or no. This can be represented in two ways such as “True” and “False” or 1 and 0. Binary data is used heavily for classification machine learning models. Examples of binary variables can include whether a person has stopped their subscription service or not, or if a person bought a car or not.

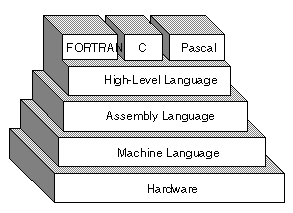
****

**1.3 Introduction to High Level Programming Language + Integrated Development Environment(IDE)**

**1.3.1 Introduction to High Level Programming Language**

Programming Languages contains set of instructions that produce various kinds of output. Programming languages are used to implement algorithms in computer programming.

Programming languages are used to give commands to computing device.



Programming languages are of three types:

1. Machine Language
2. Assembly Language
3. High Level Language
4. **Machine Languge**

Machine Language is the language contain strings of binary 1’s and 0’s. It is the only language which a computer understands without translating program.

It is easy for machine to understand it but difficult for human being to read and understand.

Disadvantage –

1. It is machine dependent (it differs from computer to computer)
2. Difficult to program and write
3. It is prone to error
4. It is difficult to modify
5. **Assembly Language**

It is low level programming language that allows a user to write a program using alphanumeric codes, instead of numeric codes for a set of instructions.

It require a translator known as assembler to convert assembly language into machine language so that it can be understood by the computer. It is easier to remember and write than machine language.

Advantage –

1. It is easy to understand and use
2. It is easy to correct and locate errors.
3. It is easy to modify

Disadvantage –

1. It is machine dependent

1. **High Level Language**

It is machine independent language. It enables users to write programs in English words and mathematical symbols. COBOL was the first high level language developed for business purpose. The examples of high-level languages are:

* Fortran
* COBOL
* Basic
* Pascal
* [C](https://codescracker.com/c/index.htm)
* [C++](https://codescracker.com/cpp/index.htm)
* [Java](https://codescracker.com/java/index.htm)

**Compiler** : A compiler is a translator which translate a high level programming language into equivalent machine language programs. It compiles a set of machine language instructions for every high level language program.

The compiler translate source code into machine level language known as object code which can be saved and executed by the user.

**Interpreter :** An interpreter is a translator used for translating high level language into desired output. It takes one statement at a time, translate it into machine language instructions and then immediately executes the result.

Advantage :

1. It is machine independent
2. It is easier to learn and use
3. It is easier to maintain and gives few errors

Disadvantage :

1. It lowers efficiency
2. It is less flexible

High-level languages are basically symbolic languages that use English words and/or mathematical symbols rather than mnemonic codes. Each instruction in the high-level language is translated into many machine language instructions that the computer can understand.

Types of high level languages are

**1) Algebraic Formula-Type Processing**

These languages are oriented towards the computational procedures for solving mathematical and statistical problems.

Examples include:

* BASIC (Beginners All Purpose Symbolic Instruction Code)
* FORTRAN (Formula Translation)
* PL/I (Programming Language, Version 1)
* ALGOL (Algorithmic Language)
* APL (A Programming Language)

**2. Business Data Processing**

These languages are best able to maintain data processing procedures and problems involved in handling files. Some examples include:

* COBOL (Common Business Oriented Language)
* RPG (Report Program Generator)

**3. String and List Processing**

These are used for string manipulation, including search patterns and inserting and deleting characters. Examples are:

* LISP (List Processing)
* Prolog (Program in Logic)

**4. Object-Oriented Programming Language**

In OOP, the computer program is divided into objects. Examples are:

* C++
* Java

**5. Visual Programming Language**

These programming languages are designed for building Windows-based applications.Examples are:

* Visual Basic
* Visual Java
* Visual C

**1.3.2 Integrated Development Environment(IDE)**

An integrated development environment (IDE) is a software suite that combines basic tools required to write and test software.

An IDE contains code editor, a compiler or interpreter, and a debugger, accessed through a single [graphical user interface](https://searchwindevelopment.techtarget.com/definition/GUI) (GUI).

The user writes and edits [source code](https://searchmicroservices.techtarget.com/definition/source-code) in the code editor. The compiler translates the source code into a readable language that is executable for a computer. And the debugger tests the software to solve any issues or bugs.

Some IDE are open source and other are commercial.

Nowdays, instead of writing code for java in notepad we use Netbeans, Eclipse or Java Editor for it.

**Benefits of IDE:**

1. Because of faster setup IDE improves productivity of software.
2. Simplifies various tasks of developer by supporting of various tools and environment.
3. Standardize the development process by organizing the necessary features for software development in the UI.

**1.4 Exploratory Data Analysis(EDA) + Data Visualization**

**1.4.1 Exploratory Data Analysis(EDA)**

Exploratory Data Analysis(EDA) is an approach used to analyse data sets to summarize data set characteristics using visual methods. EDA and data Visualization both mainly targets to represent data in graphical format.

EDA involves the analyst trying to get a “feel” for the data set, often using their own judgment to determine what the most important elements in the data set are. For example, [multidimensional scaling](https://www.statisticshowto.datasciencecentral.com/multidimensional-scaling/) is an EDA that uses visual representations of distances or similarities between sets of objects; It’s up to the user to interpret exactly what the distances represent.

Purpose of EDA

The purpose of exploratory data analysis is to:

* Check for missing data and other mistakes.
* Gain maximum insight into the data set and its underlying structure.
* Uncover a [parsimonious model](https://www.statisticshowto.datasciencecentral.com/parsimonious-model/), one which explains the data with a minimum number of [predictor variables](https://www.statisticshowto.datasciencecentral.com/independent-variable-definition/).
* Check assumptions associated with any model fitting or [hypothesis test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/).
* Create a list of [outliers](https://www.statisticshowto.datasciencecentral.com/find-outliers/)or other anomalies.
* Find [parameter](https://www.statisticshowto.datasciencecentral.com/what-is-a-parameter-statisticshowto/)estimates and their associated [confidence intervals](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/confidence-interval/)or [margins of error](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/margin-of-error/).
* Identify the most influential [variables](https://www.statisticshowto.datasciencecentral.com/variable/).

Types of Exploratory Data Analysis

EDA falls into four main areas:

* Univariate non-graphical — looking at one variable of interest, like age, height, income level etc.
* Univariate graphical.
* Multivariate non-graphical — analysis of multiple variables at the same time.
* Multivariate graphical.

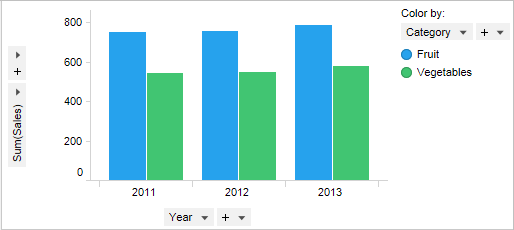
**1.4.2 Data Visualization**

Data visualization is the communication of data in a visual manner. Data visualization is the process of displaying data or information in graphical charts, figures and bars.

Visualization techniques are :

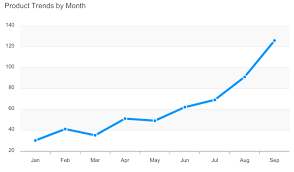
1. Bar Chart
2. Line Chart
3. Scatterplot
4. Pie Chart
5. Histogram
6. Box Plot
7. **Bar Chart**

A **bar chart** or **bar** graph is a **chart** or graph that presents categorical data with rectangular **bars** with heights or lengths proportional to the values that they represent. The **bars** can be plotted vertically or horizontally. A vertical **bar chart** is sometimes called a line graph.

****

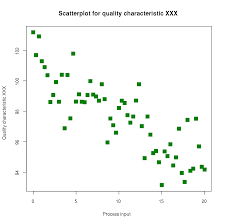
1. **Line Chart**

A **line chart** or **line** plot or **line** graph is a type of **chart** which displays information as a series of data points called 'markers' connected by straight **line** segments. It is a basic type of **chart** common in many fields.

****

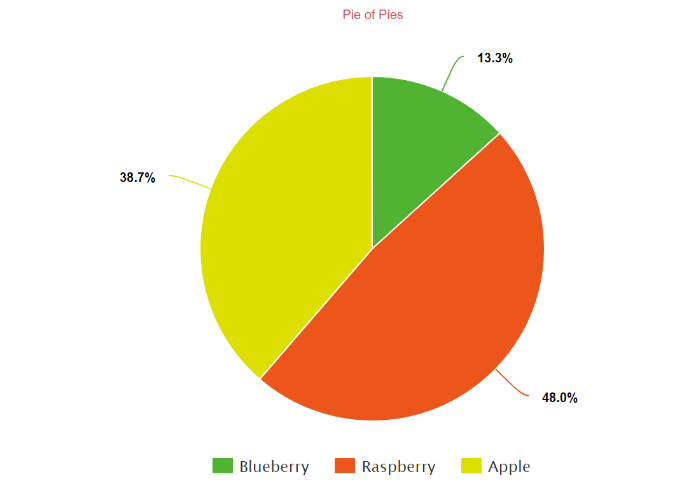
1. **Scatterplot**

A **scatter plot** is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis. For example this**scatter plot** shows the height and weight of a fictitious set of children.

****

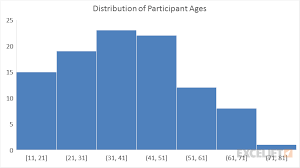
1. **Pie Chart**

A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice, is proportional to the quantity it represents.

****

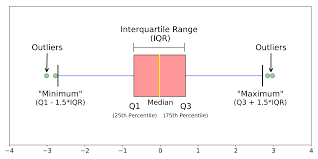
1. **Histogram**

A **Histogram** visualises the distribution of data over a continuous interval or certain time period. Each bar in a**histogram** represents the tabulated frequency at each interval/bin.

****

1. **Box Plot**

A **box and whisker plot**—also called a **box plot**—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a **box plot**, we draw a **box** from the first quartile to the third quartile. A vertical line goes through the**box** at the median.

****

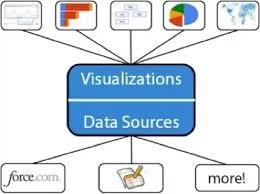
**Benefits of Data Visulization**

1. Quick, clear understanding of the information.
2. Identify emerging trends and act quickly based on what we see**.**
3. Identify relationships and patterns within digital assets
4. Analysis at various levels of detail**.**
5. Can share our story with others**.**

**1.5 Different Types of Data Sources**

A **data source** is simply the **source** of the **data** which is available for data analysis in decision-making process. It can be a file, a particular database on a DBMS, or even a live **data** feed. Data might be located locally or available remotely.

Data source is the connection set up to a database from a server.



There are various data sources which are available which includes,

1. **File System :** File system controls how data is stored and retrieved. E.g. CSV files, excel files etc.
2. **Relational System** : A **relational** database refers to a database that stores data in a structured format, using rows and columns. E.g. Oracle, Sql server, DB2 etc.
3. **Cloud System** : A **cloud system** or **cloud** computing technology refers to the computing components (hardware, software and infrastructure) that enable the delivery of **cloud** computing services such as: SaaS (software as a service), PaaS (platform as a service) and IaaS (infrastructure as service) via a network (i.e. the Internet). E.g. Amazon EC2, Apple iCloud
4. **Websites** : A **website** or web site is a collection of related network web resources, such as web pages, multimedia content, which are typically identified with a common domain name, and published on at least one web server.
5. **Live Streaming Data** : **Streaming data** is **data** that is continuously generated by different sources. ... **Data streaming** can also be explained as a technology used to deliver content to devices over the internet, and it allows users to access the content immediately, rather than having to wait for it to be downloaded.
6. **Electronic Sensors** : An **electronic sensor** is a device designs to indicate or measure something about the environment which it is operating. This information or data is created mechanically or electrically and transmits the environment information by creating an electrical signal.