

# Data Science



**By:**  
**Dr. Shikha Deep**

# Data Science

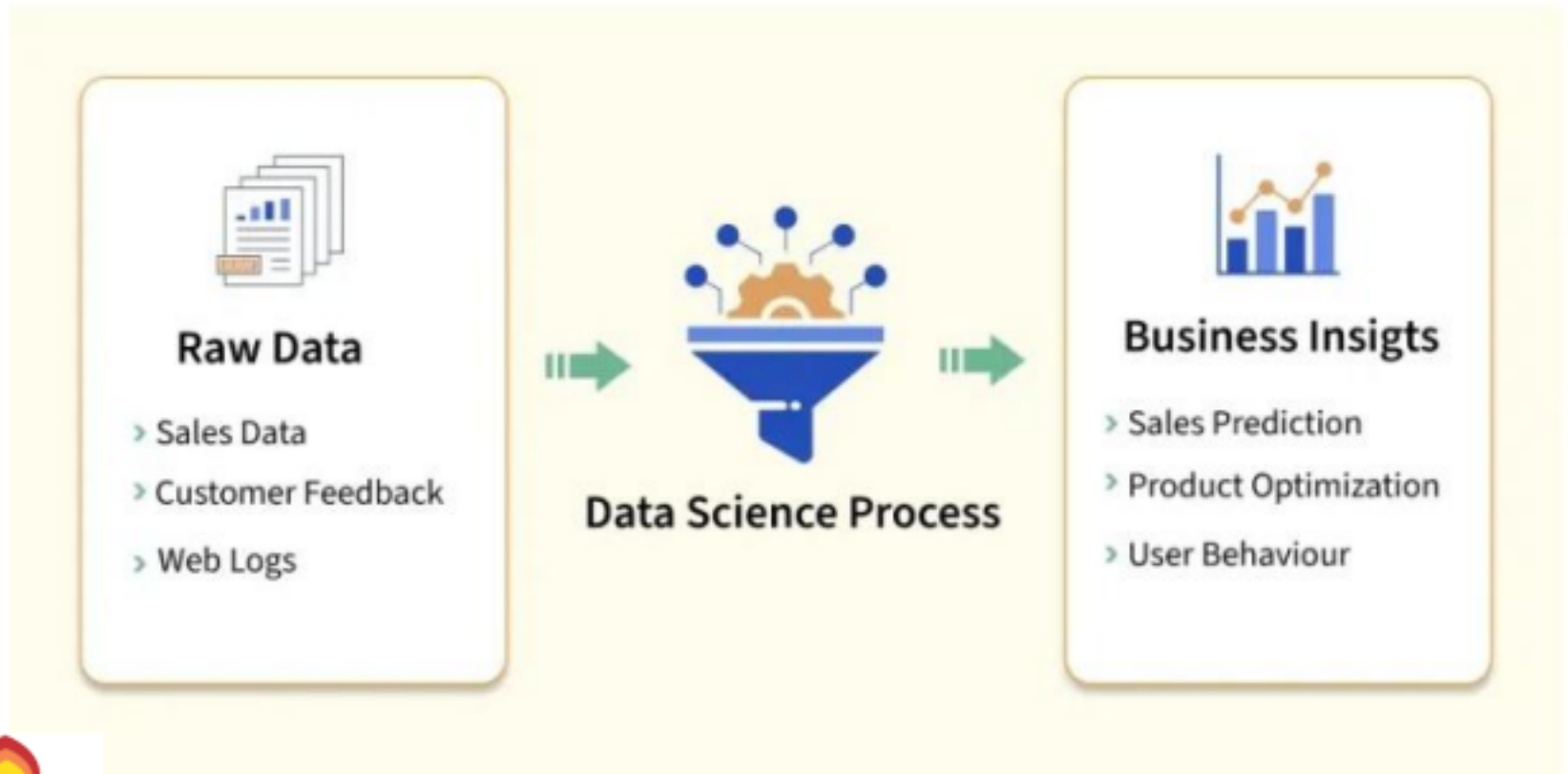
- **Data science** is a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
  - *VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.*
- **Data Science** is the process of collecting, analyzing, and using data to make smart decisions. It combines programming, statistics, and domain knowledge to find hidden patterns and trends in both structured (like Excel tables) and unstructured data (like images, videos, text).

Data Science processes the raw data and solve business problems and even make prediction about the future trend or requirement.

**For example,**

From the huge raw data of a company, data science can help answer following question:

- a) What do customer want?
- b) How can we improve our services?
- c) What will the upcoming trend in sales?
- d) How much stock they need for upcoming festival.



# Key Components of Data Science

<b>1</b>	<b>Data Collection</b>	Gather raw data from files, APIs, web, sensors, databases
<b>2</b>	<b>Data Cleaning</b>	Fix errors, missing values, wrong types, duplicates
<b>3</b>	<b>Exploratory Data Analysis (EDA) and Visualization</b>	Understand data through statistics and visualizations
<b>4</b>	<b>Feature Engineering</b>	Create, modify, and select the most useful input features
<b>5</b>	<b>Modeling (ML)</b>	Apply machine learning algorithms to learn and predict
<b>6</b>	<b>Decision Making / Interpretation</b>	Analyze model results and communicate insights

# Data Collection

- Data collection is the process of gathering and measuring information on targeted variables to answer relevant questions.
  - *Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.*
- **Data Collection** is the process of gathering raw information from various sources such as databases, sensors, or user interactions to be used for analysis, modeling, and insights.

# Resources for Data Collection

<b>Manual Entry</b>	Surveys, Google Forms, Feedback Forms
<b>APIs</b>	Twitter API, Weather API, YouTube Data API
<b>Web Scraping</b>	Extracting news from websites, job listings
<b>IoT Devices</b>	Sensors in smart homes, health trackers
<b>Files</b>	CSV, Excel, JSON, XML files
<b>Online Datasets</b>	Kaggle, UCI Repository, Government portals

# Data Cleaning

- Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset. -

Kelleher, J. D., & Tierney, B. (2018). Data Science. MIT Press.

- It is also called **Data Preprocessing**.
- It means ensuring the data is **accurate, complete, and ready** for analysis.



# Important?







- Machine Learning models require clean, consistent data
- Unclean data leads to wrong predictions or misleading insights
- Preprocessing ensures quality input = reliable output



Unitedworld Institute of Technology

## Common Problems in Raw Data

--	--

 Missing values	Empty cells in an Excel or CSV file
 Duplicates	Same record repeated
 Inconsistent formats	<b>Male, male, MALE</b>
 Outliers	A student with <b>height = 400 cm</b>
 Wrong data types	Age stored as text like <b>"twenty"</b>
 Unbalanced data	90% positive, 10% negative classes



Unitedworld Institute of Technology

# Python Methods

Check nulls	<code>df.isnull().sum()</code>	Check missing data in dataset
Remove missing rows	<code>df.dropna()</code>	Clean rows with missing values
Fill missing values	<code>df.fillna(value)</code>	Fill blanks with mean, 0, etc.
Remove duplicates	<code>df.drop_duplicates()</code>	Ensure each row is unique
Change data type	<code>df.astype(type)</code>	Fix number stored as text
Standardize text	<code>str.lower()</code> , <code>strip()</code>	Clean and format string columns
Remove outliers	z-score, IQR	Eliminate extreme values
Encode categories	<code>pd.get_dummies()</code> , <code>LabelEncoder()</code>	Use for machine learning



# Data Analysis & Visualization

- EDA is the process of analyzing data sets to summarize their main characteristics, often with visual methods. -

*Tukey, J. W. (1977). **Exploratory Data Analysis.***

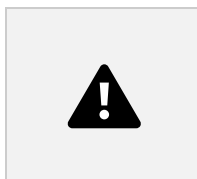
*Addison-Wesley.*

- **EDA** is the process of understanding your dataset through summaries and visualizations before diving into modeling or decision-making.
- Apply statistical and computational methods, to identify patterns, trends, or relationship of the data.



Unitedworld Institute of Technology • During visualization of data the following patterns we can observed:

❑ The following bar chart showing average gene expression for five genes (e.g., BRCA1, TP53, EGFR, KRAS, PTEN) in two conditions (Healthy vs. Cancer).



Unitedworld Institute of Technology

Histogram



Unitedworld Institute of Technology

**Why is it  
Important?**

It helps us to understand our data before modeling, like:

(a) What's the shape and distribution of the data?

(b) Are there any missing values or outliers? (c) How are different variables related?



Unitedworld Institute of Technology

## Key Steps in EDA

--	--	--	--

<b>1</b>	Data Inspection	Check structure, size, types, and basic stats of data (head(), info(), describe())	pandas
<b>2</b>	Detect Missing/Invalid Data	Find missing values and data type issues to decide on cleaning methods	pandas, seaborn
<b>3</b>	Examine Numerical Features	Analyze distributions, variance, and outliers using plots	pandas, matplotlib, seaborn
<b>4</b>	Explore Categorical Data	Study frequency of each category and class imbalance	pandas, seaborn
<b>5</b>	Find Relationships	Identify correlation or dependency among variables (e.g., salary vs age)	pandas, seaborn
<b>6</b>	Identify Outliers & Patterns	Detect extreme or rare values and discover trends	pandas, matplotlib, seaborn





Unitedworld Institute of Technology

# For Visualization of Data

<b>histplot()</b>	Show distribution of a numeric column	seaborn
<b>boxplot()</b>	Detect outliers	seaborn
<b>heatmap()</b>	Show correlation between features	seaborn
<b>countplot()</b>	Frequency of categories	seaborn
<b>scatterplot()</b>	Show relationship between two variables	seaborn



Unitedworld Institute of Technology

# Difference Between Data Cleaning and EDA?

<b>Goal</b>	Fix errors and prepare data for analysis/modeling	Understand data, discover patterns, trends, and insights
<b>Focus</b>	Correcting data	Exploring and summarizing data visually/statistically
<b>Tasks</b>	Remove nulls, fix types, handle outliers	Plot distributions, check relationships, visualize data
<b>Tools Used</b>	pandas, sklearn.preprocessing, numpy	pandas, matplotlib, seaborn, plotly

<b>Comes First?</b>	Data Cleaning is often done before and during EDA	EDA is done after initial cleaning and iteratively
<b>Output</b>	Clean dataset, ready for ML	Key insights, summary reports, visuals



Unitedworld Institute of Technology

## Featuring Engineering

- Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models.
- *Géron, A. (2019). Hands-On Machine Learning with Scikit Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.*
- **Feature Engineering** means creating new columns or

modifying existing ones so that machine learning models can understand patterns better.

For example, converting "Gender" into 0 and 1, or extracting year from a DOB.



Unitedworld Institute of Technology

## Types of Feature Engineering Tasks

<b>Feature Creation</b>	Create new columns from existing data	Combine date and time into a timestamp, extract year from DOB
<b>Feature Transformation</b>	Change scale, format, or distribution	Normalize age
<b>Feature Selection</b>	Choose only relevant columns	Remove columns with low variance

<b>Feature Encoding</b>	Convert categorical to numeric	Convert Gender = ['Male', 'Female'] to 0/1
<b>Feature Scaling</b>	Bring all numeric features to same scale	Convert marks from 0–100 to 0.0–1.0 using MinMaxScaler



Unitedworld Institute of Technology

# Modeling (ML Algorithm)

- Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. - *Mitchell, T. M. (1997). Machine Learning. McGraw-Hill*
- **Machine Learning** means teaching computers to learn

patterns from data and use smart algorithms to make predictions — like:

(a) predicting house prices.

(b) detecting spam emails or classifying emails as spam or

(c) recommending movies on Netflix.



Unitedworld Institute of Technology

# Types

<b>Supervised Learning</b>	Learn from labeled data (input → output)	Email → Spam/Not Spam, Hours → Marks
<b>Unsupervised Learning</b>	Discover patterns in unlabeled data	Grouping customers, Image compression

<b>Reinforcement Learning</b>	Learn by reward/punishment through trial & error	Game playing AI, Self driving cars
-----------------------------------	--	---------------------------------------



Unitedworld Institute of Technology

# Decision-Making

- The final step in a data science project is to interpret the results and translate them into actionable decisions. -

*Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly Media.*

- Once the model is ready, we use its results to help in real-life decision-making — like predicting demand, reducing risk, or suggesting what product to show a



user.

Unitedworld Institute of Technology

# Basic Syntax of Python

- Syntax in Python refers to the set of rules that define how a Python program is written and interpreted.
- *Zelle, J. (2016). Python Programming: An Introduction to Computer Science (3rd ed.). Franklin, Beedle & Associates.*
- **Python syntax** is the grammar of the language — the way we write Python commands so the computer understands and runs them without errors.



Unitedworld Institute of Technology



# Basic Syntax

1.	<b>Variables</b>	<code>x = 10</code>	Used to store values
2.	<b>Data Types</b>	<code>a = "Hello" (string), b = 4.5 (float)</code>	Different types: int, float, str
3.	<b>Comments</b>	<code># This is a comment</code>	Ignored by Python, used to explain code
4.	<b>Print</b>	<code>print("Welcome")</code>	Displays output on the screen
5.	<b>Input</b>	<code>name = input("Enter name: ")</code>	Takes user input from the console
6.	<b>List</b>	<code>marks = [85, 90, 95]</code>	A collection of items
7.	<b>If Condition</b>	<code>if x &gt; 0: print("Positive")</code>	Makes decisions based on logic
8.	<b>Loop (for)</b>	<code>for i in range(3): print(i)</code>	Repeats a block of code

**9. Function** `def add(x, y): return x + y` A reusable block of code with a name



**Unitedworld Institute of Technology**

