KARNAVATI UNIVERSITY

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,

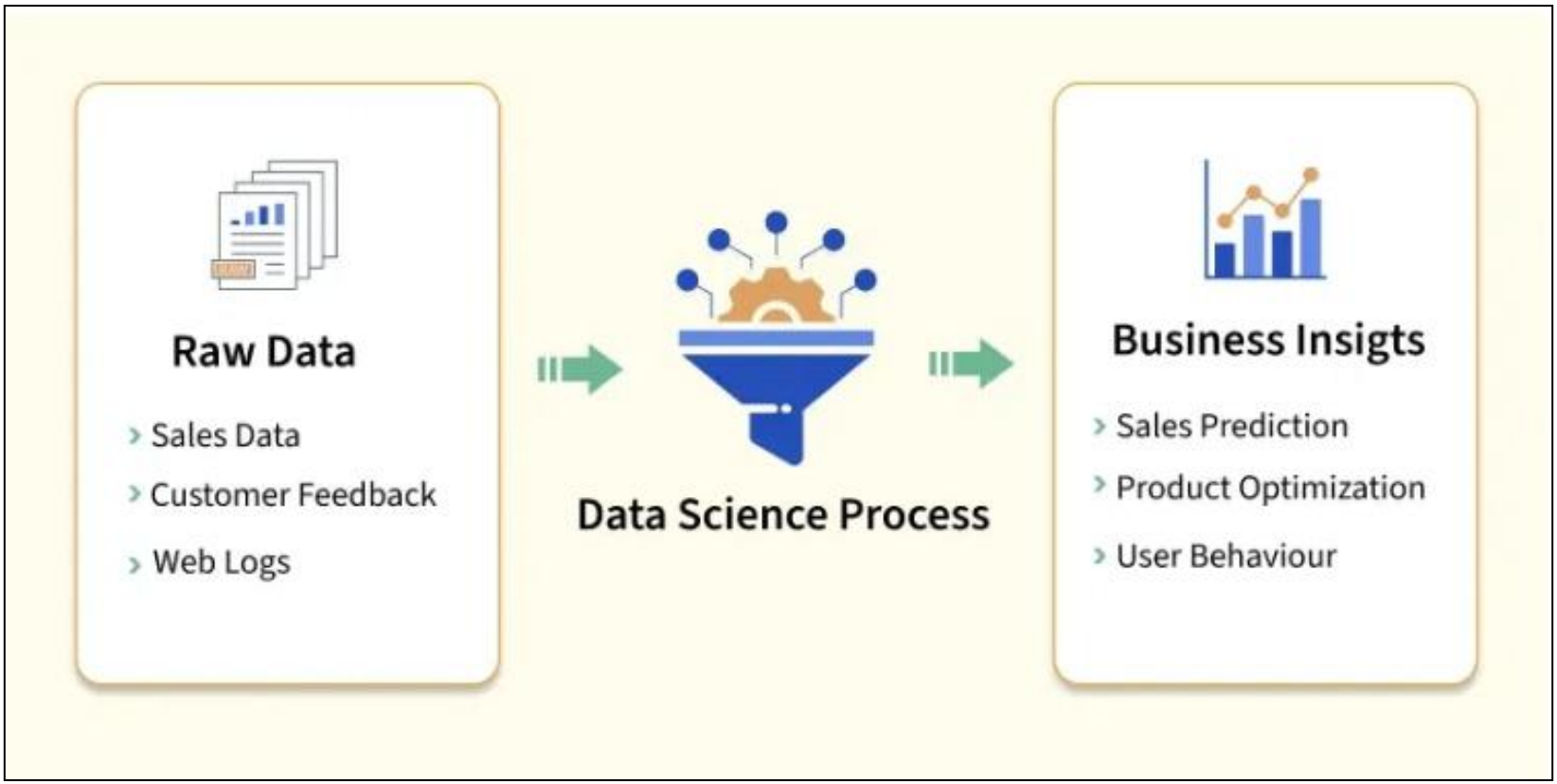
From the huge raw data of a company, data science can help answer following question:

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.


Raw Data
Sales Data
Customer Feedback
Web Logs
Data Science Process
Business Insigts
Sales Prediction
Product Optimization
User Behaviour

Key Components of Data Science

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

Source Type	Example
Manual Entry	Surveys, Google Forms, Feedback Forms
APIs	Twitter API, Weather API, YouTube Data API
Web Scraping	Extracting news from websites, job listings
IoT Devices	Sensors in smart homes, health trackers
Files	CSV, Excel, JSON, XML files
Online Datasets	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.

Why is it Important?

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

Common Problems in Raw Data

Problem	Example
❌ Missing values	Empty cells in an Excel or CSV file
📄 Duplicates	Same record repeated
🧹 Inconsistent formats	**Male, male, MALE**
📏 Outliers	A student with **height = 400 cm**
🔤 Wrong data types	Age stored as text like **"twenty"**
⚖️ Unbalanced data	90% positive, 10% negative classes

Python Methods

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

Data Analysis & Visualization

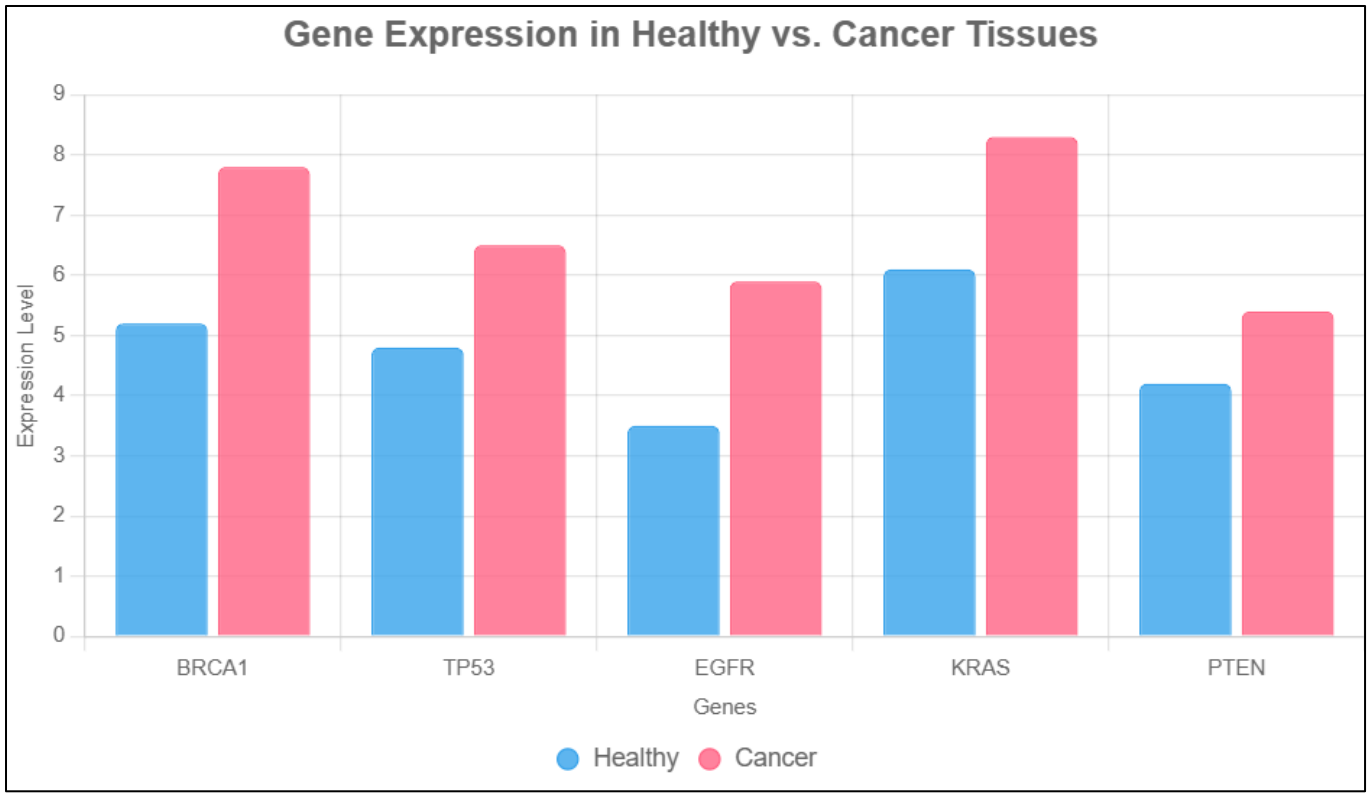
- EDA is the process of analyzing data sets to summarize their main characteristics, often with visual methods.

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

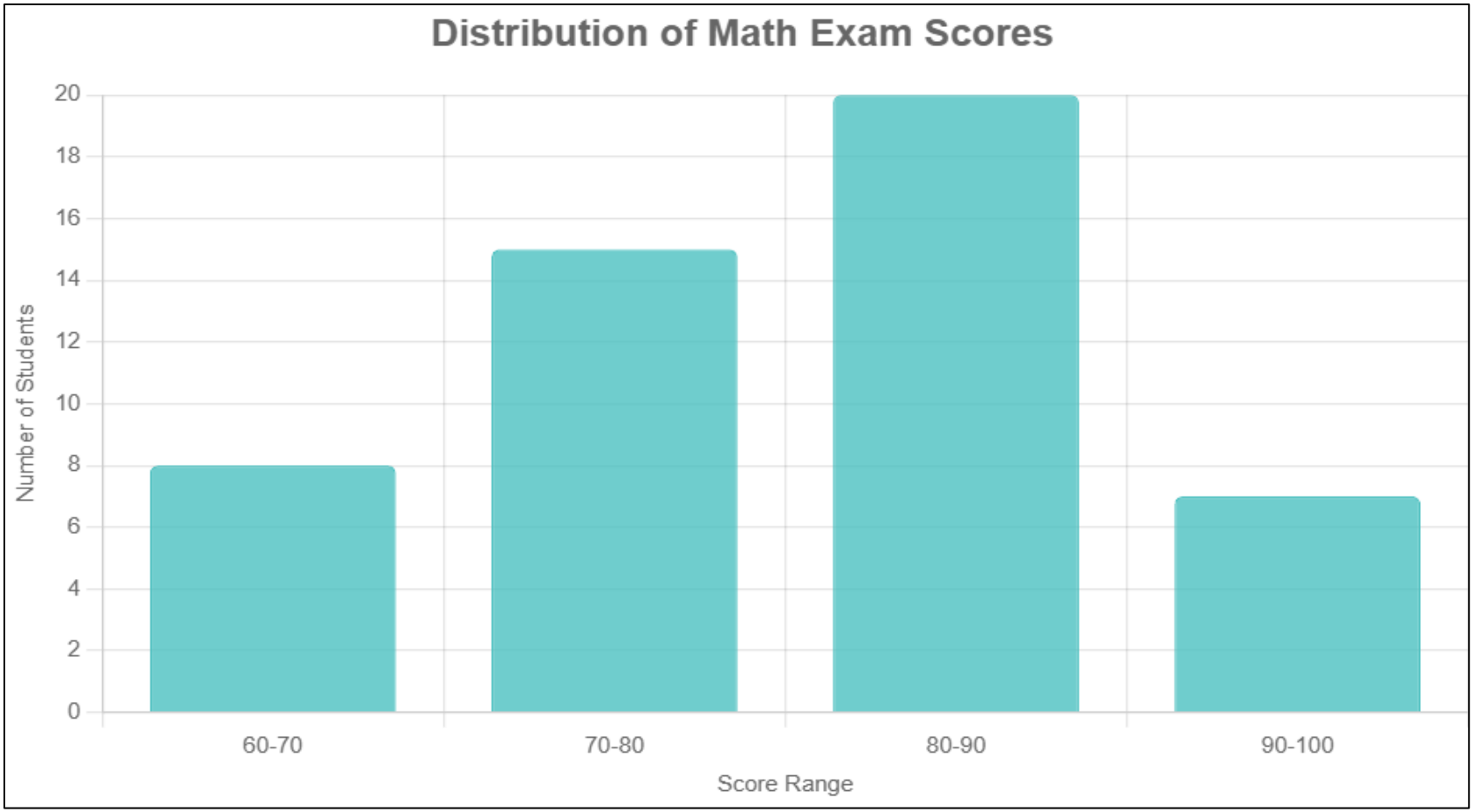
- **EDA** is the process of understanding your dataset through summaries and visualizations before diving into modeling or decision-making.
- Apply statistical and computational methods, to identify patterns, trends, or relationship of the data.

- 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).



Histogram


Distribution of Math Exam Scores
Number of Students
0
2
4
6
8
10
12
14
16
18
20
60-70
70-80
80-90
90-100
Score Range

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?

Key Steps in EDA

S.No.	EDA Feature	One-Line Explanation	Common Libraries
1	Data Inspection	Check structure, size, types, and basic stats of data (head(), info(), describe())	pandas
2	Detect Missing/Invalid Data	Find missing values and data type issues to decide on cleaning methods	pandas, seaborn
3	Examine Numerical Features	Analyze distributions, variance, and outliers using plots	pandas, matplotlib, seaborn
4	Explore Categorical Data	Study frequency of each category and class imbalance	pandas, seaborn
5	Find Relationships	Identify correlation or dependency among variables (e.g., salary vs age)	pandas, seaborn
6	Identify Outliers & Patterns	Detect extreme or rare values and discover trends	pandas, matplotlib, seaborn

For Visualization of Data

Plot Type	Purpose	Library Used
histplot()	Show distribution of a numeric column	seaborn
boxplot()	Detect outliers	seaborn
heatmap()	Show correlation between features	seaborn
countplot()	Frequency of categories	seaborn
scatterplot()	Show relationship between two variables	seaborn

Difference Between Data Cleaning and EDA?

Concept	Data Cleaning	EDA (Exploratory Data Analysis)
Goal	Fix errors and prepare data for analysis/modeling	Understand data, discover patterns, trends, and insights
Focus	Correcting data	Exploring and summarizing data visually/statistically
Tasks	Remove nulls, fix types, handle outliers	Plot distributions, check relationships, visualize data
Tools Used	pandas, sklearn.preprocessing, numpy	pandas, matplotlib, seaborn, plotly
Comes First?	Data Cleaning is often done before and during EDA	EDA is done after initial cleaning and iteratively
Output	Clean dataset, ready for ML	Key insights, summary reports, visuals

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.

Types of Feature Engineering Tasks

Task Type	Description	Example
Feature Creation	Create new columns from existing data	Combine date and time into a timestamp, extract year from DOB
Feature Transformation	Change scale, format, or distribution	Normalize age
Feature Selection	Choose only relevant columns	Remove columns with low variance
Feature Encoding	Convert categorical to numeric	Convert Gender = ['Male', 'Female'] to 0/1
Feature Scaling	Bring all numeric features to same scale	Convert marks from 0–100 to 0.0–1.0 using MinMaxScaler

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 not,

(c) recommending movies on Netflix.

Types

Type	Description	Examples
Supervised Learning	Learn from labeled data (input → output)	Email → Spam/Not Spam, Hours → Marks
Unsupervised Learning	Discover patterns in unlabeled data	Grouping customers, Image compression
Reinforcement Learning	Learn by reward/punishment through trial & error	Game playing AI, Self-driving cars

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.

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.

Basic Syntax

S.No.	Topic	Example	Explanation
1.	**Variables**	x = 10	Used to store values
2.	**Data Types**	a = "Hello" (string), b = 4.5 (float)	Different types: int, float, str
3.	**Comments**	# This is a comment	Ignored by Python, used to explain code
4.	**Print**	print("Welcome")	Displays output on the screen
5.	**Input**	name = input("Enter name: ")	Takes user input from the console
6.	**List**	marks = [85, 90, 95]	A collection of items
7.	**If Condition**	if x > 0: print("Positive")	Makes decisions based on logic
8.	**Loop (for)**	for i in range(3): print(i)	Repeats a block of code
9.	**Function**	def add(x, y): return x + y	A reusable block of code with a name


Thanks!