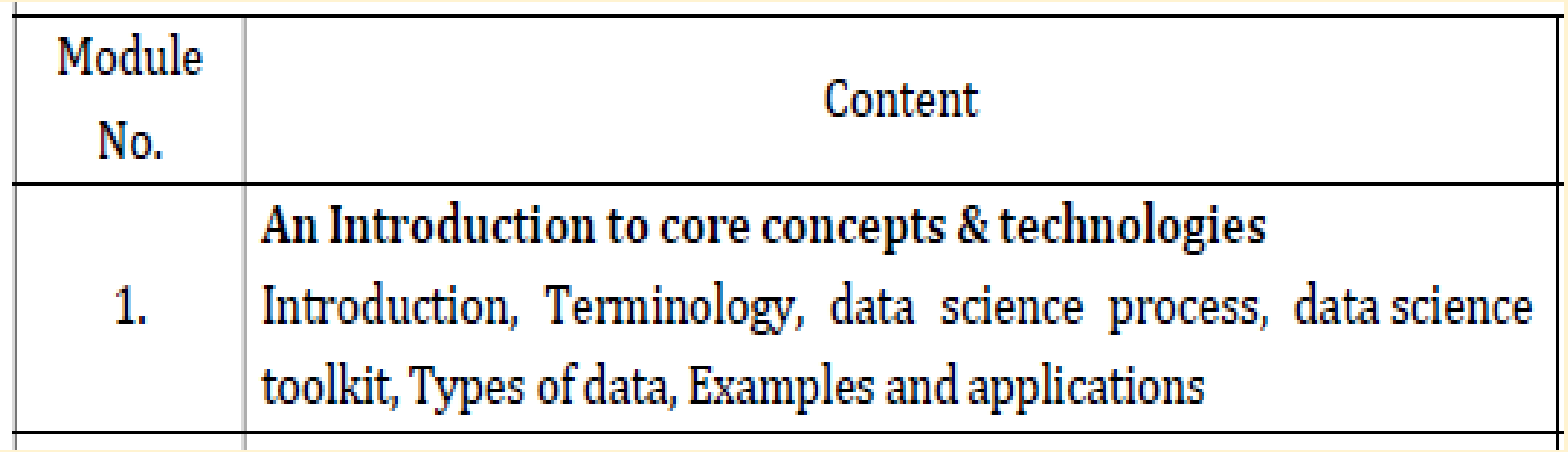
DATA SCIENCE-SSCA3021

MODULE-I

Prepared by- Mr. Anurag Anand

MODULE-I



Module No.	Content
1.	**An Introduction to core concepts & technologies** Introduction, Terminology, data science process, data science toolkit, Types of data, Examples and applications

1. Introduction

- **What is Data Science?**

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- It combines elements of statistics, computer science, and domain expertise.

Why is Data Science Important?

- **Data Explosion:** We live in an era where massive amounts of data are generated every second. Data science provides the tools to make sense of this deluge.
- **Informed Decision Making:** Businesses and organizations can make data-driven decisions rather than relying on intuition.
- **Predictive Power:** It allows for forecasting future trends and behaviors, which is crucial for planning and strategy.
- **Innovation:** It drives new products, services, and efficiencies across various industries.

Examples of Data Science in Action:

- **Netflix Recommendations:** Suggesting movies and TV shows based on your viewing history and preferences.
- **Google Search:** Ranking search results based on relevance.
- **Fraud Detection:** Identifying unusual patterns in financial transactions to flag potential fraud.
- **Self-Driving Cars:** Using sensor data to perceive the environment and navigate.

2. Terminology

- **Data:** Raw facts, figures, or values. It can be qualitative (descriptive) or quantitative (numerical).

Example: A customer's age (numerical), gender (categorical), and feedback comments (textual).

- **Information:** Data that has been processed, organized, and structured in a way that makes it meaningful and useful.

Example: An average age of customers, or a summary of common themes in customer feedback.

2. Terminology

- **Knowledge:** The understanding of relationships and patterns derived from information, leading to insights and actionable conclusions.

Example: Discovering that customers between 25-34 years old who gave positive feedback often mentioned "ease of use" of a product.

- **Insight:** A deep understanding of a person or thing, often derived from data analysis, that reveals hidden truths or motivations.

Example: Identifying that a specific marketing campaign resonated particularly well with a certain demographic, leading to a higher conversion rate.

2. Terminology

- **Model:** A mathematical representation of a real-world process or relationship, learned from data. Models are used for prediction, classification, or understanding.

- *Example:* A linear regression model predicting house prices based on features like size and number of bedrooms.

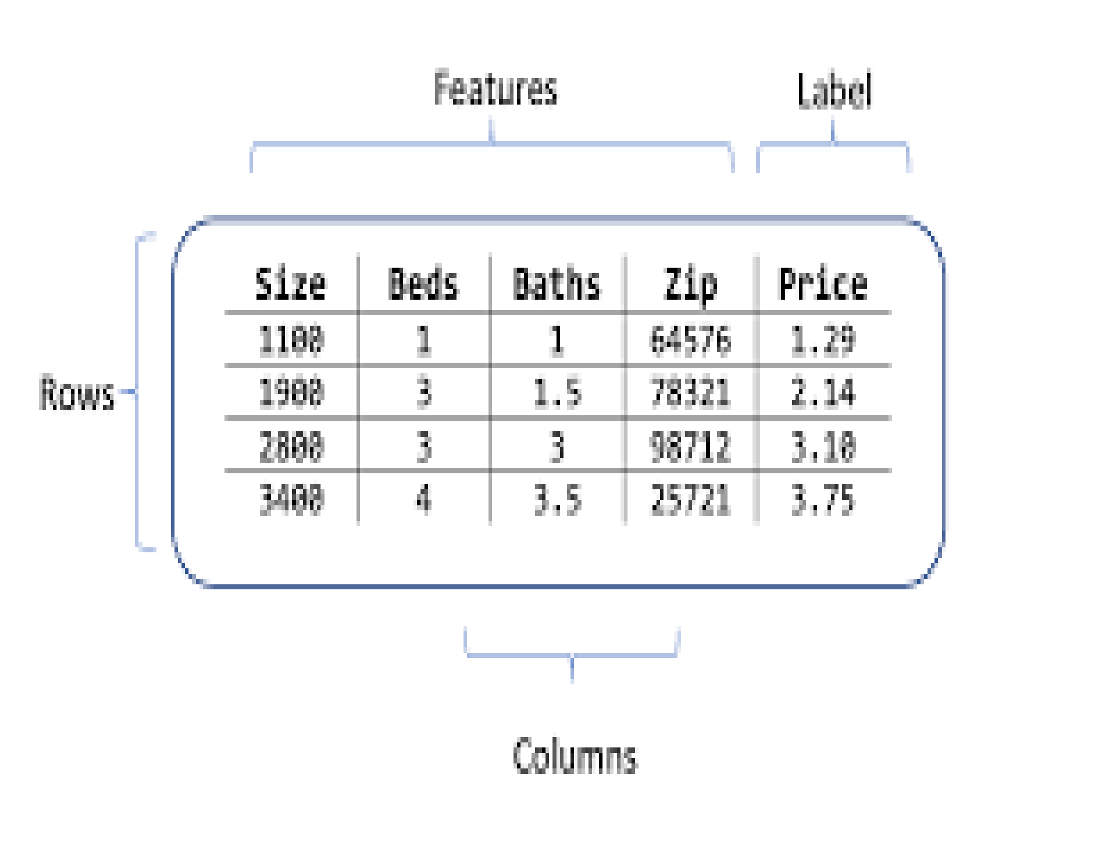
- **Algorithm:** A set of well-defined instructions or rules to solve a problem or perform a computation. In data science, algorithms are used to build models.

- *Example:* The k-Nearest Neighbors (k-NN) algorithm for classification.

2. Terminology

- **Features/Variables:** The individual measurable properties or characteristics of the data that are used as input to a model.
- *Example:* In predicting house prices, features could be "square footage," "number of bedrooms," "location."

- **Target/Label:** The variable that an algorithm is trying to predict or classify.
- *Example:* In house price prediction, the "price" is the target. In classifying emails as spam or not, "spam" or "not spam" is the label.



Size	Beds	Baths	Zip	Price
1100	1	1	64576	1.29
1900	3	1.5	78321	2.14
2800	3	3	98712	3.10
3400	4	3.5	25721	3.75

3. Data Science Process

- This outlines the typical steps involved in a data science project. While the exact terminology might vary, the core phases remain consistent. A common framework is the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology.

3. Data Science Process

- **1. Business Understanding/Problem Definition:**

- **Description:** Clearly defining the problem you're trying to solve, the objectives, and the expected outcomes from a business perspective. This involves understanding the domain and what success looks like.
- *Example:* A retail company wants to reduce customer churn. The objective is to identify customers at high risk of churning and implement targeted retention strategies.

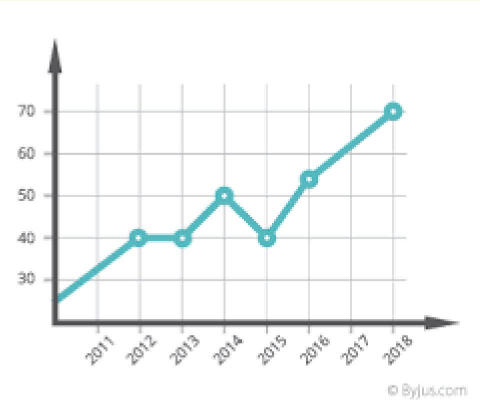
- **2. Data Acquisition/Collection:**

- **Description:** Identifying and gathering the necessary data from various sources (databases, APIs, web scraping, logs, etc.).
- *Example:* Collecting customer transaction history, demographic information, website interaction data, and customer service call logs.

3. Data Science Process

- **3. Data Cleaning/Preparation (Data Wrangling):**
- **Description:** Transforming raw data into a clean, consistent, and suitable format for analysis. This is often the most time-consuming step. It includes handling missing values, outliers, inconsistencies, and formatting issues.
- *Example:* Removing duplicate customer records, correcting misspelled city names, filling in missing age values using imputation, and converting categorical data into numerical representations.

3. Data Science Process



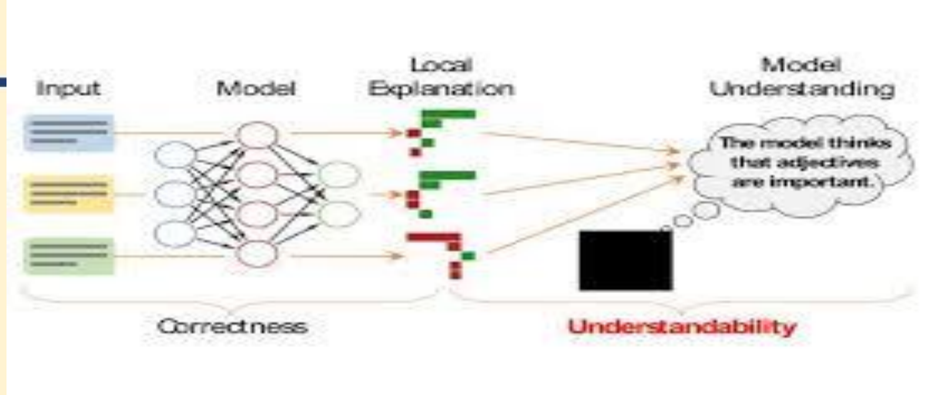
4 Exploratory Data Analysis (EDA):

- **Description:** Analyzing data sets to summarize their main characteristics, often with visual methods. EDA helps in understanding data patterns, detecting anomalies, and formulating hypotheses.
- *Example:* Plotting the distribution of customer ages, visualizing the correlation between purchase frequency and churn, or identifying trends in product returns.

5. Feature Engineering:

- **Description:** Creating new features from existing ones to improve the performance of machine learning models. This often requires domain knowledge.
- *Example:* From transaction data, creating a new feature like "average monthly spend" or "number of unique products purchased in the last 6 months."

3. Data Science Process



- **6. Model Building/Selection:**
- **Description:** Choosing and applying appropriate machine learning algorithms to build a predictive or descriptive model based on the prepared data. This involves training the model on a portion of the data.
- *Example:* Using a logistic regression model or a decision tree to predict customer churn.

- **7. Model Evaluation/Validation:**
- **Description:** Assessing the performance of the built model using various metrics and testing it on unseen data (validation set) to ensure its generalization ability.
- *Example:* Measuring the accuracy, precision, recall, or F1-score of the churn prediction model on a test dataset.

3. Data Science Process

8. Deployment/Production:

- **Description:** Integrating the validated model into a live system or application so it can be used to make predictions or generate insights in real-time or batch.
- *Example:* Integrating the churn prediction model into the CRM system to automatically flag high-risk customers for customer service representatives.

9. Monitoring and Maintenance:

- **Description:** Continuously monitoring the model's performance in a production environment and retraining it periodically with new data as data patterns evolve.
- *Example:* Tracking the actual churn rates against the predicted churn rates and retraining the model quarterly to adapt to changing customer behavior.

4. Data Science Toolkit

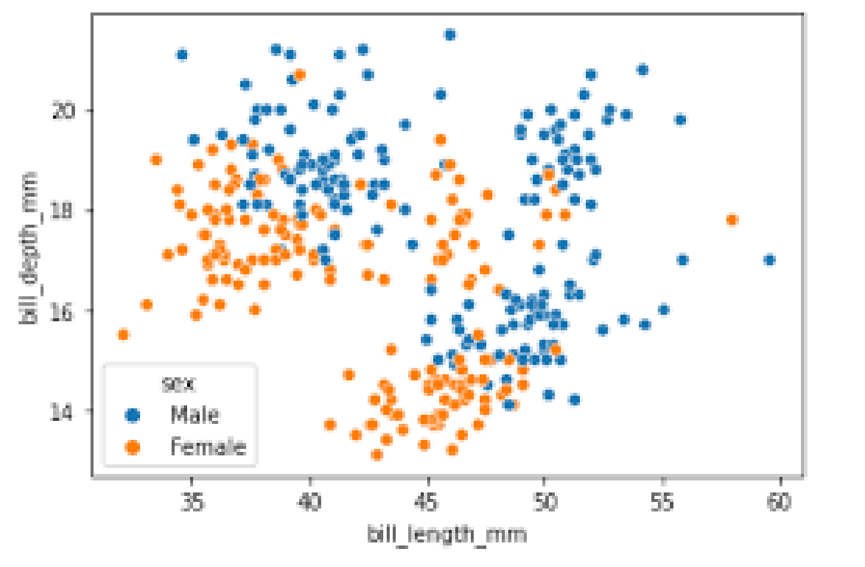
- **<u>Programming Languages:</u>**
  - **Python:** Widely used due to its extensive libraries for data manipulation, analysis, and machine learning.
    - *Example Libraries:* Pandas (data manipulation), NumPy (numerical computing), Scikit-learn (machine learning), Matplotlib/Seaborn (visualization).
  - **R:** Popular among statisticians for statistical analysis and graphical representation.
    - *Example Packages:* dplyr (data manipulation), ggplot2 (visualization), caret (machine learning).
  - **SQL (Structured Query Language):** Essential for interacting with and extracting data from relational databases.
    - *Example:* SELECT * FROM Customers WHERE Age > 30;

4. Data Science Toolkit

Data Manipulation and Analysis Libraries/Tools:

- **Pandas (Python):** For data structures like DataFrames (similar to spreadsheets) and operations like filtering, grouping, merging.
  - *Example:* df['Sales'].mean() to calculate the average sales
  -
- **NumPy (Python):** For numerical operations, especially with arrays.
  - *Example:* Performing element-wise operations on large datasets.

4. Data Science Toolkit



- **<u>Visualization Tools:</u>**
- **Matplotlib (Python):** Basic plotting library
- **Seaborn (Python):** Built on Matplotlib, offers more aesthetically pleasing statistical plots.
- **Plotly (Python/R/JavaScript):** For interactive visualizations.
- **Tableau/Power BI:** Business intelligence tools for creating interactive dashboards.
  - *Example:* Creating a bar chart to show sales by region in Tableau

4. Data Science Toolkit

Machine Learning Frameworks:

- **Scikit-learn (Python):** Comprehensive library for various ML algorithms (classification, regression, clustering).
  - *Example:* from sklearn.linear_model import LogisticRegression; model = LogisticRegression().fit(X_train, y_train)
- **TensorFlow/Keras (Python):** For deep learning (neural networks).
- **PyTorch (Python):** Another popular deep learning framework.

4. Data Science Toolkit

- **Big Data Technologies (for large datasets):**
- **Apache Hadoop:** Distributed storage and processing of large datasets.
- **Apache Spark:** In-memory distributed processing engine, faster than Hadoop.
- **NoSQL Databases (e.g., MongoDB, Cassandra):** For unstructured or semi-structured data.

4. Data Science Toolkit

- **Cloud Platforms:**
- **AWS (Amazon Web Services):** Offers services like S3 (storage), EC2 (compute), SageMaker (ML platform).
- **Google Cloud Platform (GCP):** Offers services like BigQuery (data warehouse), AI Platform.
- **Microsoft Azure:** Offers services like Azure Blob Storage, Azure Machine Learning.

5. Types of Data

Understanding different data types is crucial because it influences the appropriate analysis techniques and models.

- **1. Numerical Data (Quantitative Data):** Represents measurable quantities.
  - **a. Discrete Data:** Can only take specific, distinct values (often integers).
    - *Example:* Number of children in a family (you can't have 2.5 children), number of cars in a parking lot.
  - **b. Continuous Data:** Can take any value within a given range.
    - *Example:* Height of a person (1.75m, 1.755m, etc.), temperature (25.3°C), time.

5. Types of Data

2. Categorical Data (Qualitative Data): Represents qualities or characteristics that cannot be measured numerically.

a. Nominal Data: Categories with no intrinsic order or ranking.

- *Example:* Colors (Red, Blue, Green), Gender (Male, Female, Non-binary), Marital Status (Single, Married, Divorced).

b. Ordinal Data: Categories with a clear order or ranking, but the intervals between categories are not necessarily equal.

- *Example:* Education Level (High School, Bachelor's, Master's, PhD), Customer Satisfaction (Very Dissatisfied, Dissatisfied, Neutral, Satisfied, Very Satisfied), T-shirt sizes (S, M, L, XL).

5. Types of Data

3. Time Series Data: A sequence of data points indexed in time order.
Example: Stock prices over days, hourly temperature readings, monthly sales figures.

4. Text Data: Unstructured or semi-structured data consisting of human language.
Example: Customer reviews, emails, social media posts, news articles.

5. Image/Audio/Video Data: Complex unstructured data types.
Example: Medical images (X-rays), voice recordings for speech recognition, surveillance video footage.

6. Examples and Applications

- **Healthcare:**

- **Application:** Disease prediction (e.g., predicting diabetes risk based on patient data), drug discovery, personalized medicine, medical image analysis (detecting tumors in X-rays/MRIs).

- *Example:* Using machine learning to analyze patient demographics, lab results, and medical history to identify individuals at high risk of developing heart disease, allowing for early intervention.

6. Examples and Applications

- **Finance:**

- **Application:** Fraud detection (credit card fraud, insurance fraud), algorithmic trading, credit scoring, risk assessment.

- *Example:* A bank uses a data science model to analyze transaction patterns and identify unusual activities that might indicate fraudulent credit card use, blocking suspicious transactions in real-time

6. Examples and Applications

- **E-commerce/Retail:**

- **Application:** Recommendation systems (products, movies), customer segmentation, demand forecasting, pricing optimization, churn prediction.

- *Example:* Amazon uses data science to recommend products based on your Browse history, past purchases, and items viewed by similar customers, leading to increased sales.

6. Examples and Applications

- **Marketing:**
- **Application:** Targeted advertising, campaign optimization, customer lifetime value prediction, sentiment analysis of brand mentions.
- *Example:* A marketing team uses data science to identify which customer segments respond best to certain types of ads, optimizing their budget and improving campaign ROI.

6. Examples and Applications

- **Transportation/Logistics:**
- **Application:** Route optimization, traffic prediction, autonomous vehicles, fleet management.
- *Example:* Google Maps uses real-time and historical traffic data, along with machine learning, to predict traffic congestion and suggest the fastest routes