---

1. Introduction to Data Science (Beginner)

- What is Data Science?
    - Definition and importance
    - History and evolution
- Data Science vs. Related Fields
    - Data science vs. statistics, data analysis, and machine learning
- Data Science Workflow
    - Problem definition
    - Data collection
    - Data cleaning
    - Exploratory data analysis (EDA)
    - Modelling
    - Model evaluation and deployment
- Roles in Data Science
    - Data scientist, data analyst, data engineer, machine learning engineer

---

2. Data Types and Structures (Beginner)

- Data Types
    - Structured vs. unstructured data
    - Qualitative vs. quantitative data
- Basic Data Structures
    - Arrays, lists, dictionaries (Python)
    - Data frames (Pandas)
    - Relational databases (SQL)

---

3. Data Collection and Data Sources (Beginner)

- Methods of Data Collection

    - Surveys, web scraping, APIs
- Types of Data
    - Primary data (directly collected)
    - Secondary data (external datasets)
- Data Storage
    - SQL databases
    - NoSQL databases (MongoDB)

---

4. Data Cleaning and Preprocessing (Beginner-Intermediate)

- Handling Missing Data
    - Imputation, removal, or replacement
- Handling Outliers
    - Detection and treatment
- Normalization and Standardization
    - Scaling data (min-max, Z-score)
- Feature Engineering
    - Creating new features from existing ones
- Data Transformation
    - Encoding categorical variables
    - Log transformation for skewed data

---

5. Exploratory Data Analysis (EDA) (Beginner-Intermediate)

- Descriptive Statistics
    - Mean, median, mode, standard deviation, variance
- Data Visualization
    - Histograms, scatter plots, box plots, heatmaps
    - Tools: Matplotlib, Seaborn, Polly
- Correlation and Covariance

    - Understanding relationships between variables

---

6. Introduction to Probability and Statistics (Intermediate)

- Probability Basics
    - Probability theory, probability distributions (normal, binomial, Poisson)
- Descriptive vs. Inferential Statistics
    - Mean, median, standard deviation vs. hypothesis testing
- Hypothesis Testing
    - Null vs. alternative hypotheses, p-value, confidence intervals
- Sampling Techniques
    - Random sampling, stratified sampling, sampling bias

---

7. Data Visualization (Intermediate)

- Principles of Data Visualization
    - Communicating insights effectively
    - Choosing the right visualization type
- Data Visualization Tools
    - Python: Matplotlib, Seaborn, Polly
    - BI Tools: Power BI, Tableau
    - Advanced visualizations: Dashboards, interactive charts

---

8. Introduction to Machine Learning (Intermediate)

- Supervised Learning
    - Regression (Linear, Polynomial)
    - Classification (Logistic regression, Decision trees, K-Nearest Neighbours)
- Unsupervised Learning
    - Clustering (K-Means, DBSCAN)
    - Dimensionality Reduction (PCA, t-SNE)

- Model Evaluation Metrics
    - Accuracy, precision, recall, F1-score, confusion matrix
- Overfitting and Underfitting
    - Regularization techniques (L1, L2)

---

9. Advanced Machine Learning (Advanced)

- Ensemble Learning
    - Random Forest, Gradient Boosting, Boost
- Support Vector Machines (SVM)
- Neural Networks and Deep Learning
    - Basic neural networks
    - Convolutional neural networks (CNN)
    - Recurrent neural networks (RNN)
    - Frameworks: TensorFlow, Kera's, Porch
- Natural Language Processing (NLP)
    - Text preprocessing (tokenization, stemming, lemmatization)
    - Sentiment analysis, text classification
    - Topic modelling (Latent Dirichlet Allocation)

---

10. Feature Engineering and Selection (Intermediate-Advanced)

- Feature Selection Techniques
    - Recursive feature elimination, correlation-based methods
- Dimensionality Reduction
    - Principal Component Analysis (PCA)
    - Feature importance (decision trees, Random Forests)

---

11. Model Tuning and Optimization (Advanced)

- Hyperparameter Tuning

    - Grid search, Random search
- Cross-Validation
    - K-fold cross-validation
    - Train-test split
- Model Evaluation
    - ROC-AUC, precision-recall curves
    - Bias-variance trade-off

---

12. Big Data and Distributed Computing (Advanced)

- Introduction to Big Data
    - Definition and characteristics (volume, velocity, variety)
    - Tools and technologies: Hadoop, Spark
- MapReduce
    - Basic concepts and operations
- Cloud Computing for Data Science
    - AWS, Google Cloud, Azure for data storage and processing

---

13. Data Ethics and Governance (Intermediate)

- Data Privacy and Security
    - GDPR, data anonymization
- Bias in Data and Models
    - Ethical considerations, fairness in algorithms
- Responsible AI: Ensuring ethical usage of AI

---

14. Time Series Analysis (Intermediate-Advanced)

- Introduction to Time Series Data
    - Trend, seasonality, noise
- Time Series Models

    - ARIMA, SARIMA, Exponential Smoothing
- Advanced Time Series Models
    - LSTM (Long Short-Term Memory) networks for time-series forecasting

---

15. Data Science in Practice (Case Studies and Applications) (All Levels)

- Industry Use Cases
    - Healthcare (predictive modelling, disease outbreak forecasting)
    - Finance (fraud detection, credit scoring)
    - Retail (customer segmentation, recommendation systems)
    - Marketing (targeted advertising, A/B testing)

---

16. Capstone Projects and Real-World Implementation (Advanced)

- End-to-End Data Science Projects
    - From data collection to model deployment
    - Tools for deployment: Flask, Docker, Kubernetes
- Working with Stakeholders
    - Presenting insights, business impact

---

17. Advanced Topics and Trends (Cutting-Edge Topics)

- Artificial Intelligence (AI) and Deep Learning
    - Advanced neural networks (GANs, autoencoders)
    - Transfer learning
- Reinforcement Learning
    - Markov Decision Processes, Q-learning
- Explainable AI (XAI)
    - Understanding and interpreting AI decisions
- Quantum Computing for Data Science
    - Theoretical foundations and applications in data science

WHAT IS DATA SCIENCE

A field that works with data to understand patterns, make decision and solve problems

Its helps because it helps the company make informed decision

History and Evolution of Data Science

Data science has been evolved from traditional statistics, which focused on collecting and analyzing data to a broader field that include big data, machine learning and artificial intelligence

Components of data Science

Data Collection: Gathering data from various sources (surveys, online databases etc.)

Data Processing and Cleaning: Organizing and fixed the data to ensure its accurate and usable

Data Analysis and Interpretation: looking for trends and pattern in the data

Data Visualization and Communication: Creating charts or graphs to share findings in an understandable way

DATA SCIENCE WORKFLOW

Problem Definition: Clearly stating the question and what you want to do

Data Acquisition: Finding and collecting the necessary data

Data Exploration: Analyzing the data to understand its structure and quality

Model Building: Creating a model (a simplified representation of reality) to predict outcomes based on data

Model Evaluation and Deployment: Testing the model to see how well it works and using it in real life scenarios

TYPES OF DATA

Structured Data: Organized data, like spreadsheet (egg customer name and sales figure)

Unstructured Data: Organized data, like emails or social media (egg, post, text)

DATA SOURCES:

Internal data: data collected within organization

External Data: Data obtained from outside sources (egg public, datasets, data from other companies)

APIs and Web Scraping: Method to gather data from website or online services

KEY TOOLS AND TECHNOLOGIES

Programming Languages

Python and R are used for statistics

Tools

Jupiter Notebook: An environment to write and run code interactively

RStudio: A tool specifically for R programming

Power BI: A tool for creating interactive reports and dashboards

Databases:

SQL: Languages manage and retrieve data from databases

NoSQL: A type of database for unstructured data

Roles in Data Science

Data Scientist: Analyze data to uncover insight and build model

Data Analyst: Focuses on interpreting data and creating reports

Data Engineer: Build and maintain data pipelines to ensure data is accessible

Machine learning Engineer: Develop algorithm that allows computer to learn from data

Challenges in data science

Data Quality Issue: poor data can lead to incorrect conclusion

Ethical consideration: Ensuring Data Privacy and avoiding bias in data analysis

Scalability and Performance: Handling large volume of data efficiently

Data Collection

Survey: collecting data directly through peoples

Web Scraping: using software tools to extract data from website

APIs (Application Programming Interfaces): Accessing data from external application or platforms (egg: Twitter APIs to get tweets)

Databases: Gathering data from store databases (SQL) or NoSQL

Types of data

Primary data: data collected from direct resources (surveys)

Secondary data: data collected from resources but are reused (online dataset, government reports)

Data Processing and Cleaning

Why it is important: so that we do not process incorrect data

- Common data cleaning task
- Handling missing values
- Removing duplicates
- Correcting Errors
- Normalization

Data Analysis and Interpretation

- Exploratory data analysis (EDA)
- First step you take to get to know your data better, before you dive into more complex data analyses or modelling
- It helps you understand the general structure of your data and spot patterns and issue
- Techniques that can be used
    1. Summarization: min, max, average
    2. Visualization: pie charts, bar charts, Scatter plot etc.