Course Syllabus: Synthetic Data Engineer

Course Title: Crafting Data That Isn't Real: Mastering Synthetic Data for AI

Target Audience: Suitable for data scientists, engineers, analysts, and students interested in creating artificial datasets for AI training. Basic knowledge of Python and data concepts is helpful but not required.

Course Level: Comprehensive program covering Basic, Intermediate, and Advanced levels.

Duration: 12 weeks (flexible for self-paced learning).

Course Description:
This course trains students to become Synthetic Data Engineers, experts in generating and managing artificial datasets to train AI models when real data is scarce, sensitive, or biased. You'll learn to create realistic synthetic data, ensure its quality, and use it to power AI systems, like those for Zomato's billing or recommendation platforms. From basic data generation to advanced privacy-preserving techniques, you'll build skills to support secure, ethical AI development.

Learning Objectives:

Upon completion, students will be able to:

- Understand the role of synthetic data in AI training and testing.
- Generate synthetic datasets using statistical and AI-based methods.
- Evaluate the quality and utility of synthetic data for AI models.
- Apply privacy-preserving techniques (e.g., differential privacy) to synthetic data.
- Use synthetic data in real-world AI applications (e.g., restaurant analytics).
- Develop a portfolio of synthetic data projects.

Course Structure:

Part 1: Basic Foundations (Weeks 1-4)

This section introduces synthetic data and its role in AI.

- Week 1: Introduction to Synthetic Data
    - What is synthetic data? Why use it (privacy, cost, availability)?
    - Role of a Synthetic Data Engineer.
    - Examples: Synthetic customer data for Zomato's billing tests.
    - Exercise: Explore a synthetic dataset sample.
- Week 2: Data Basics and Tools
    - Data types: Tabular, time-series, images.
    - Python tools: pandas, NumPy, Faker for data generation.
    - Hands-on: Generate a basic synthetic dataset (e.g., fake restaurant orders).
- Week 3: Statistical Data Generation
    - Statistical methods: Sampling, distributions (e.g., Gaussian).
    - Preserving data patterns: Mean, variance, correlations.
    - Exercise: Create synthetic data matching a real dataset's statistics.
- Week 4: Evaluating Synthetic Data
    - Metrics: Statistical similarity, utility for AI training.
    - Tools: SDMetrics, Synthetic Data Vault.
    - Hands-on Project: Generate and evaluate synthetic data for a Zomato-like billing system.

Part 2: Intermediate Concepts (Weeks 5-8)

This section focuses on AI-based data generation and quality assurance.

- Week 5: AI-Based Synthetic Data
    - Generative models: GANs, VAEs for synthetic data.
    - Generating tabular data with CTGAN.
    - Hands-on: Use a GAN to create synthetic customer data.

- Week 6: Synthetic Data for Specific Use Cases
    - Generating time-series data (e.g., order timestamps).
    - Synthetic data for testing AI models (e.g., recommendation systems).
    - Case Study: Synthetic data for Zomato's delivery predictions.
- Week 7: Privacy and Security
    - Introduction to differential privacy.
    - Anonymizing sensitive data (e.g., customer info).
    - Hands-on: Apply differential privacy to a synthetic dataset.
- Week 8: Quality Assurance and Validation
    - Testing synthetic data: Utility, fidelity, privacy checks.
    - Comparing synthetic vs. real data performance.
    - Hands-on Project: Create and validate synthetic data for an AI model.

Part 3: Advanced & Expert-Level Application (Weeks 9-12)

This section prepares students for enterprise-grade synthetic data solutions.

- Week 9: Advanced Generative Models
    - Diffusion models for high-quality synthetic data.
    - Multimodal data: Combining text, images (e.g., menus, orders).
    - Exercise: Generate multimodal synthetic data for a food app.
- Week 10: Scaling Synthetic Data Pipelines
    - Building automated data generation pipelines.
    - Tools: Apache Airflow, AWS for data workflows.
    - Hands-on: Design a scalable synthetic data pipeline.
- Week 11: Ethics and Compliance
    - Ethical issues: Bias in synthetic data, misuse risks.
    - Compliance: GDPR, CCPA for synthetic data.
    - Exercise: Audit synthetic data for ethical compliance.

- Week 12: Capstone Project & Trends
    - Capstone Project: Develop a synthetic dataset for a Zomato-like platform, supporting an AI model (e.g., billing or recommendations).
    - Trends: Synthetic data for federated learning, edge AI.
    - Career paths: Data engineering, AI privacy, consulting.

Assignments & Grading:

- Weekly Data Labs & Exercises: 25%
- Intermediate Projects (Weeks 4 & 8): 30%
- Capstone Project: 35%
- Class Participation & Peer Reviews: 10%