INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
Dundigal - 500 043, Hyderabad, Telangana

COURSE CONTENT

EXPLORATORY DATA ANALYSIS								
VIII Semester: CSE(DS)								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
		L	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
ACDD34	Open Elective	3	0	0	3	40	60	100
Contact Classes: 48	**Tutorial Classes: Nil**	**Practical Classes: Nil**				**Total Classes: 48**		
Prerequisites: Basic understanding of statistics, probability, linear algebra, Python programming, and data visualization libraries like Matplotlib or Seaborn.								

I. COURSE OVERVIEW:

This course introduces the Exploratory Data Analysis (EDA) is a critical step in the data science pipeline used to understand data distributions, uncover relationships, detect anomalies, and generate hypotheses through visual and statistical methods. This course introduces students to essential techniques in EDA, using tools such as Python, Pandas, Matplotlib, and Seaborn. Through hands-on analysis of real-world datasets, students will learn how to summarize and visualize data effectively, enabling better model selection and insight discovery.

II. COURSE OBJECTIVES:

The students will try to learn:

I. The principles and importance of exploratory data analysis in the data science process.
II. The statistical techniques for summarizing, describing, and transforming raw datasets.
III. The methods for detecting patterns, outliers, and relationships using visualization tools.
IV. The practical implementation of EDA using Python libraries and real-world datasets.

III. COURSE OUTCOMES:

CO1 **Describe** the purpose and process of exploratory data analysis in data science workflows.
CO2 **Summarize** datasets using statistical metrics, data distributions, and aggregation techniques.
CO3 **Visualize** univariate, bivariate, and multivariate relationships using charts and plots.
CO4 **Identify** missing values, outliers, skewness, and data quality issues.
CO5 **Apply** feature engineering and data transformation techniques for deeper insight.
CO6 **Perform** complete EDA on real datasets and present findings through structured reports and dashboards.

IV. COURSE CONTENT:

Module I: Introduction to EDA & Data Preparation (10)
Introduction to EDA, Role of EDA in data science, Types of data (categorical, numerical, ordinal), Importing and cleaning data, Handling missing values, Data types and conversions, Identifying and correcting inconsistent data, Data subsetting and filtering, Basic summary statistics, Working with Pandas Data Frames.

Module II: Univariate Analysis (10)
Distribution of numerical variables, Histograms and density plots, Box plots and violin plots, Descriptive statistics (mean, median, mode, variance), Outlier detection (IQR, Z-score), Frequency tables for categorical variables, Bar plots and pie charts, Skewness and kurtosis, Handling skewed distributions, Binning and discretization.

Module III: Bivariate and Multivariate Analysis (10)
Scatter plots and correlation, Pair plots and heatmaps, Cross-tabulation and pivot tables, Group-wise analysis.

ANOVA and t-tests for comparison, Categorical vs numerical visualizations, Multivariate box plots, Interaction effects, Matrix plots, Color and size encoding in plots.

Module IV: Data Transformation and Feature Engineering (10)
Log and square-root transformation, Scaling and normalization, Label encoding and one-hot encoding, Feature extraction from dates, Creating new features, Dimensionality reduction overview, PCA introduction, Treating outliers via transformation, Mapping and conditional logic in features, Domain-driven feature engineering.

Module V: EDA Project and Communication (8)
EDA project pipeline overview, Choosing the right plots, Creating dashboards with Seaborn/Plotly, Data storytelling with visualizations, Structuring EDA reports, Case study: EDA on real-world datasets (e.g., Titanic, Retail, Healthcare), Presenting insights to stakeholders, Best practices and ethical use of data.

V. TEXTBOOKS:

1. Tukey, John W., *Exploratory Data Analysis*, Addison-Wesley, 1977.
2. Ben Jones, *Communicating Data With Tableau*, O'Reilly Media, 2014.
3. Hadley Wickham & Garrett Grolemund, *R for Data Science*, O'Reilly, 2017 (Chapters on EDA applicable).
4. Jake VanderPlas, *Python Data Science Handbook*, O'Reilly Media, 2016.

VI. REFERENCE BOOKS:

1. **Allen B. Downey**, *Think Stats: Exploratory Data Analysis in Python*, O'Reilly Media, 2nd Edition, 2014.
2. **Claus O. Wilke**, *Fundamentals of Data Visualization*, O'Reilly Media, 2019.
3. **Nathan Yau**, *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*, Wiley, 2011.

VII. ELECTRONICS RESOURCES:

1. NPTEL – Data Science for Engineers (EDA Modules Included)
   https://nptel.ac.in/courses/106106179

2. Kaggle – EDA Notebooks and Competitions
   https://www.kaggle.com/learn/pandas
3. Coursera – Applied Data Science with Python (UMich)
   https://www.coursera.org/specializations/data-science-python
4. Real Python – EDA with Pandas and Matplotlib
   https://realpython.com/pandas-python-explore-dataset/

VIII. MATERIALS ONLINE

1. Course template
2. Tutorial question bank
3. Tech talk topics
4. Open-ended experiments
5. Definitions and terminology
6. Assignments
7. Model question paper – I
8. Model question paper – II
9. Lecture notes
10. PowerPoint presentation
11. E-Learning Readiness Videos (ELRV)