PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100 Feet Ring Road, BSK III Stage, Bengaluru-560 085
Department of Computer Science and Engineering
Session Aug-Dec 2025

UE24MA242A: Mathematics for Computer Science Engineers

Datathon Information and Guidelines

Datathon Objective

To analyse a given dataset and perform **Exploratory Data Analysis (EDA)**.
EDA helps answer questions such as:

- How to ensure data is ready for applying machine learning algorithms?
- How to choose suitable algorithms for a dataset?
- How to identify and define the most relevant feature variables?

Exploratory Data Analysis (EDA) is the **first step** in the data analysis process. It involves summarizing, visualizing, and understanding the dataset's main characteristics to identify patterns, trends, anomalies, or relationships.

Key Goals of EDA
EDA is used to:

- Detect **mistakes, missing data**, and anomalies
- Identify the **underlying structure** of the dataset
- Determine the **most significant variables**
- Test hypotheses and model assumptions
- Establish a **parsimonious model** (minimal but effective predictors)
- Estimate parameters and confidence intervals

Value of EDA

EDA ensures the validity, interpretability, and applicability of future results. It allows data scientists to verify data quality, detect anomalies, and refine feature variables for machine learning.
By thoroughly exploring the dataset, EDA often uncovers **hidden insights** that can be crucial for decision-making.

Tools and Techniques

Popular tools and libraries for EDA include:
Python: NumPy, SciPy, Matplotlib, Pandas, Scikit-Learn, Seaborn, BeautifulSoup
R: Various statistical and visualization packages
Sample Case Studies

- House Prices: Advanced Regression Techniques – Kaggle
- Exploratory Data Analysis – Retail Case Study

Data Sources

- Kaggle Datasets

Evaluation Criteria

1. Data Set understanding

- How well the dataset meets the given criteria
- Number of attributes (columns) and rows (tuples)
- Contribution and relevance of attributes
- Handling of missing values (NaNs)
- Presence of categorical and numerical features

2. Data Cleaning

- Identification and handling of missing data
- Appropriate replacement methods (categorical/numeric)
- Justification for any dropped attributes
- Use of advanced cleaning techniques as mentioned in guidelines

3. Visualization

- Appropriateness of the chosen graph (e.g., scatter plot for correlation)
- Clarity of visual representation and ease of interpretation
- Proper labeling (titles, axis names, legends)
- Correct scaling and axis extensions to avoid misinterpretation

4. Insights and Results

- Hypothesis testing and interpretation
- Data normalization and normality checks
- Graph-based conclusions
- Accuracy or performance measures

5. Key Information Nugget

- Identification of a **unique insight** or **non-obvious finding** specific to the dataset

---

Mode of conduction:

Students will be provided with a **dataset (.csv file)**. Teams of **three students** will have **4 hours** to complete the analysis and answer predefined questions.

At the end of the event, teams must submit:

- An **analysis report (PDF)**
- The **implemented code (.ipynb file or .py file)**

These files should be compressed into a **single ZIP folder**, named using the **SRNs of the team members** (e.g., SRN1_SRN2_SRN3.zip).