**Framework for Data and Visual Analytics**

**Assignment 2**

**Last date : 5/08/2025**

**Problem Statement 1: EDA on Student Academic Performance Dataset**

**Objective:**

Perform Exploratory Data Analysis (EDA) on a real-world dataset related to student academic performance to uncover insights, clean the data, and prepare it for further analysis or modeling.

**Dataset Description:**

You are provided with a dataset named student\_performance.csv, which contains academic and demographic information of students from different schools.

**Tasks to be Performed:**

1. **Load the Data:**
   * Load the CSV file into a DataFrame.
   * Check the shape and column information.
2. **Initial Inspection:**
   * View first and last 5 rows of the data.
   * Generate summary statistics using .describe().
   * Check data types of each column using .info().
3. **Data Cleaning:**
   * Detect and handle missing values in the dataset.
     + Fill missing numeric scores with the mean or median.
     + Fill missing categorical values with the mode.
   * Identify and remove any duplicate entries.
   * Strip and normalize string columns (e.g., trim spaces in names or remarks).
   * Convert all column names to lowercase and replace spaces with underscores for consistency.
4. **Filtering and Subsetting:**
   * Find students who scored above 90 in all three subjects.
   * Subset the data for students with attendance less than 75%.
   * Get the top 5 students in mathematics score from each school.
5. **Data Type Conversion:**
   * Convert attendance to float if not already.
   * Ensure student\_id is treated as a string.
   * Convert gender, school, and lunch to category type.
6. **Exploratory Analysis:**
   * Calculate average scores by gender.
   * Count how many students completed the test preparation course.
   * Visualize missing data using heatmap (if applicable).
   * Generate histograms for each numeric column.

**Problem Statement 2: EDA on E-Commerce Customer Orders**

**Objective:**

Analyze customer order data to identify purchasing trends, clean inconsistent records, and prepare the data for business insights

**Dataset Description:**

The dataset ecommerce\_orders.csv includes order details from an online retail store

**asks to be Performed:**

1. **Load and Inspect Data**:
   * Load the CSV file into a Pandas DataFrame.
   * Check for nulls, data types, and summary statistics.
2. **Data Cleaning**:
   * Handle missing values in review\_rating, payment\_method, and delivery\_status.
   * Check for and remove duplicate rows.
   * Correct inconsistent spellings in payment\_method and delivery\_status.
3. **Feature Engineering**:
   * Calculate total\_price where missing, using quantity \* unit\_price.
   * Convert order\_date to datetime format.
4. **Filtering and Subsetting**:
   * Extract orders placed in the last 6 months.
   * Filter out cancelled or incomplete orders.
   * Identify top 5 most ordered products.
5. **Data Type Conversion**:
   * Convert order\_id and customer\_id to string.
   * Convert category and payment\_method to categorical types.
6. **Exploratory Analysis**:
   * Average rating by category.
   * Distribution of payment methods.
   * Number of orders per month.

**Problem Statement 3: EDA on Vehicle Insurance Claims Dataset**

**Objective**

Explore and clean a motor insurance claims dataset to assess data quality and prepare it for potential risk analysis.

**Dataset Description:**

The dataset insurance\_claims.csv includes historical claim information:

**Tasks to be Performed:**

1. **Data Import and Exploration**:
   * Load the CSV into a DataFrame.
   * Use .info(), .head(), .describe() to explore.
2. **Data Cleaning**:
   * Identify missing values in claim\_amount, vehicle\_type, and driving\_experience.
   * Fill missing driving\_experience with median based on age groups.
   * Normalize categorical values (yes, Yes, YES → Yes).
3. **Filtering/Subsetting**:
   * Find all claims greater than ₹50,000 from SUVs.
   * Filter policies from the southern region where fraud was flagged.
   * List customers under 25 with high claim amounts.
4. **Data Type Fixes**:
   * Convert policy\_id to string.
   * Convert accident\_date to datetime format.
   * Convert region and vehicle\_type to category.
5. **Insights and Summarization**:
   * Average claim amount per vehicle type.
   * Count of fraud cases by region.
   * Histogram of driving experience.