Week 7 Lab

Due Date: Monday, 16th December 2024

Lab Session: Exploratory Data Analysis (EDA) on Heart Health Dataset

---

Objective

By the end of this lab session, students will:

1. Understand the steps involved in EDA, including loading and inspecting data, cleaning, and analyzing relationships.
2. Perform univariate, bivariate, and multivariate analysis.
3. Identify and visualize patterns, anomalies, and correlations in the dataset.

Dataset

Filename: heart_data.csv
Description: Contains information on heart health indicators, including demographic, clinical, and lifestyle variables.

Lab Tasks

Task 1: Data Loading and Inspection

1. Import necessary libraries
2. Load the dataset
3. Inspect the dataset

Task 2: Data Cleaning

1. **Handle Misrepresented Data in columns**:
    - Check each column and determine if it contains the right data
    - Encode categorical columns if necessary
2. **Handle Missing Values**:
    - Count missing values in each column
    - Decide on strategies for handling missing values
3. **Handle Duplicates**:
    - Check for duplicates

Task 3: Univariate Analysis

1. Visualize distributions of numerical variables:
    - Use histograms and boxplots
2. Analyze categorical variables:
    - Plot bar charts

Task 4: Bivariate Analysis

1. **Numerical-Numerical Relationships**:
    - Create a scatterplot to explore relationships
    - Calculate correlation matrix
2. **Categorical-Numerical Relationships**:
    - Compare target across independent variables using boxplots
3. **Categorical-Categorical Relationships**:
    - Create a crosstab to explore categories

Task 5: Multivariate Analysis

- Use pairplots to visualize multiple relationships:

Task 6: Communication of Insights

1. Summarize findings:
    - What are the most significant predictors of heart disease?
    - Are there notable patterns (e.g., higher risk in certain age groups or genders)?
2. Create visualizations to support conclusions:
    - Use PowerPoint or Jupyter Notebook to present insights with visuals.

Deliverables

1. Python script or notebook with the EDA process.
2. A report summarizing key findings (PDF or Markdown format).
3. Visuals highlighting relationships and trends.
