National University of Computer and Emerging Sciences

Lab Exercise 06
DL2001-Introduction to Data Science Lab

Course Instructor	Ms. Mariam Nasim
Lab Instructor(s)	Ms. Rida Amir
Section	BDS-3A
Semester	Fall 2025

Department of Data Science

FAST-NU, Lahore, Pakistan

Exercise

1. Life Expectancy Dataset
    a. Load the dataset using pandas.read_csv()
    b. Display the first 5 rows
    c. Use df.info() and df.describe() to get a basic understanding
    d. Check how many missing values are present per column.
    e. Drop columns with more than 30% missing values.
    f. For the remaining missing values:
        - Fill numerical columns using median.
        - Fill categorical columns (like "Status") using mode.
    g. Check the dataset size after dropping.
    h. Check if any duplicate rows exist
    i. Remove them using drop_duplicates()
    j. Check the dataset size after dropping.
    k. Identify the columns containing outliers using both IQR and Z-score method
    l. Remove these rows using both IQR and Z-score method
    m. Compare the results from both and explain which technique would be suitable and why?
    n. List all the categorical columns
    o. Apply One-Hot Encoding to all nominal categorical columns
    p. Apply Label Encoding to any ordinal categorical column (if applicable)
    q. Save the cleaned dataset as cleaned.csv and submit it with your .ipynb file

2. Diabetes Dataset
    a. Load the CSV into a pandas DataFrame.
    b. Display basic information:
        - Shape
        - Column names
        - first 5 rows
        - .info() and .describe()
    c. Plot a scatter plot color-coded by Outcome (0 vs 1) to show the relationship between:
        - Plot Glucose vs BMI,
        - Age vs Insulin
        - BloodPressure vs SkinThickness
    d. Plot box plots for Glucose, BMI, Insulin and Age separated by Outcome (0 vs 1). This helps see the distribution and detect outliers for diabetics vs non-diabetics.

e. Plot histogram for each numeric feature (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age)
f. Plot a correlation matrix (heatmap) to understand how the features are correlated.
g. Provide an analysis of the dataset based on all of the above visualizations.