TP N1: Exploratory Data Analysis (EDA)

Objective: To explore, clean, and analyse dataset using Python (pandas, Matplotlib, and seaborn).

Exercise 1: Data Loading and overview.

Objective: Understanding dataset structure.

Dataset: Titanic dataset.

Task:

1. Import necessary libraries: pandas, numpy, and matplotlib.pyplot.
2. Load the dataset into a DataFrame.
3. Display the first and last 5 rows of the dataset.
4. Show the number of rows and columns.
5. Display column names and data types.
6. Use .info() and .describe() to understand data statistics.
7. Identify categorical and numerical columns.
8. Write a short summary describing what the dataset is about and what each column represents.

Exercise 2: Data Cleaning and Preparation.

Objective: Handling missing and inconsistent data.

Dataset: Iris dataset.

Task:

1. Check for missing values using .isnull().sum().
2. Visualise missing data using a heatmap (seaborn.heatmap).
3. Decide how to handle missing data (drop or fill) and apply changes.
4. Check for duplicated rows and remove them if necessary.
5. Convert incorrect data types (e.g., change object to category or datetime).
6. Detect outliers using boxplots.
7. Apply simple techniques to handle outliers (e.g., capping or removing).
8. Summarise the cleaning steps you performed and explain why.

Exercise 3: Univariate and Bivariate Analysis.

Objective: Exploring distributions and relationships.

Dataset: Students Performance dataset.

Task:

1. Plot histograms for all numerical features.
2. Create countplots for all categorical features.
3. Identify skewness or unusual distributions in numeric data.
4. Use boxplots to compare numerical features across categories (e.g., sns.boxplot(x="gender", y="math score")).
5. Create a correlation matrix and visualise it using a heatmap.
6. Plot scatterplots to explore relationships between two numerical variables.
7. Write 3 short observations based on the plots (e.g., "Students with higher reading scores tend to have higher writing scores").

Home work:

Objective: Full EDA pipeline on a new dataset.

Dataset Suggestions:

- Netflix Movies and Shows dataset
- FIFA Players dataset
- Global Temperature Change dataset
- World Happiness dataset

Task:

1. Choose a dataset of your interest.
2. Load and explore the data (structure, summary statistics).
3. Clean the dataset: handle missing data, duplicates, and data types.
4. Conduct univariate and bivariate analysis.
5. Create at least 5 plots (histogram, boxplot, scatterplot, heatmap, barplot).
6. Write a short report summarising:
    - Dataset description
    - Cleaning steps
    - Key findings and visualisations
    - 3 final insights or recommendations