DATA SCIENCE

DL-2001

Lab 03

Python Programming

National University of Computer & Emerging Sciences – NUCES – Karachi

National University of Computer & Emerging Sciences - NUCES - Karachi

FAST School of Computing

Course Code: DL-2001	Data Science Lab

Objectives

1. Develop proficiency with NumPy for numerical computing
2. Apply Pandas for structured data handling
3. Conduct Exploratory Data Analysis (EDA)
4. Integrate multiple data science skills in problem-solving
5. Foster analytical thinking and domain understanding

NumPy for Data Science

NumPy (Numerical Python) is a fundamental library for scientific computing in Python. It provides:

- Multidimensional arrays (ndarray)
- Mathematical functions (linear algebra, statistics, etc.)
- Broadcasting and vectorization (fast computations)

```
1. import numpy as np
2.
3. # Create arrays
4. a = np.array([1, 2, 3, 4, 5])
5. b = np.array([[1, 2, 3], [4, 5, 6]])
6.
7. print("1D Array:", a)
8. print("2D Array:\n", b)
9.
```

```
1. # Arithmetic operations
2. arr1 = np.array([10, 20, 30])
3. arr2 = np.array([1, 2, 3])
4.
5. print(arr1 + arr2)   # [11, 22, 33]
6. print(arr1 * arr2)   # [10, 40, 90]
7. print(arr1 / arr2)   # [10.0, 10.0, 10.0]
8.
```

```
1. arr = np.array([1, 2, 3, 4, 5, 6])
2.
3. print("Mean:", np.mean(arr))
4. print("Median:", np.median(arr))
5. print("Standard Deviation:", np.std(arr))
6. print("Reshape:", arr.reshape(2, 3))
7.
```

Pandas

Pandas is a Python library for data analysis and manipulation. It provides two main data structures:

- Series → 1D labeled array (like a column).
- DataFrame → 2D table with labeled rows & columns.

Series in pandas

```
1. import pandas as pd
2.
3. # From a list
4. s = pd.Series([10, 20, 30, 40], index=["a", "b", "c", "d"])
5. print(s)
6.
7. # From a dictionary
8. s2 = pd.Series({"Math": 90, "English": 85, "Science": 88})
```

```
print(s2)

```

```
data = {
    "Name": ["Ali", "Sara", "John"],
    "Marks": [85, 90, 78],
    "Age": [20, 21, 19]
}

df = pd.DataFrame(data)

```

EDA

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics, discover patterns, detect anomalies, and check assumptions — often using statistics and visualization.

Importance of EDA

- Helps us understand the dataset before modeling.
- Detects missing values, duplicates, and outliers.
- Reveals relationships between variables.
- Guides feature engineering and preprocessing steps.

Steps in EDA

1. Understand data structure → rows, columns, datatypes.
2. Check for missing values & duplicates.
3. Summarize statistics (mean, median, distribution).
4. Visualize distributions (histograms, boxplots).
5. Explore relationships (scatter plots, correlation heatmap).
6. Draw insights (patterns, trends, anomalies).

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv("student_marks.csv")

# Quick look
print(df.shape)         # Rows & columns
print(df.columns)       # Column names
print(df.info())        # Data types & nulls
print(df.describe())    # Summary statistics

```

```
# Missing values
print(df.isna().sum())

# Fill missing numeric with median
df.fillna(df.median(numeric_only=True), inplace=True)

# Drop duplicate rows
df.drop_duplicates(inplace=True)

```

```
# 1.  Histogram for distribution
df["Marks"].hist(bins=10)
plt.title("Marks Distribution")
plt.show()

# Boxplot for outliers
df.boxplot(column="Marks")
plt.title("Boxplot of Marks")
plt.show()

```

```
# Scatter plot
plt.scatter(df["Study_Hours"], df["Marks"])
plt.xlabel("Study Hours")
plt.ylabel("Marks")
plt.title("Marks vs Study Hours")
plt.show()

# Group analysis
print(df.groupby("Gender")["Marks"].mean())
```

```
# Correlation matrix
corr = df.corr(numeric_only=True)
print(corr)

# Heatmap
plt.imshow(corr, cmap="coolwarm", interpolation="nearest")
plt.colorbar()
plt.xticks(range(len(corr.columns)), corr.columns, rotation=90)
plt.yticks(range(len(corr.columns)), corr.columns)
plt.title("Correlation Heatmap")
plt.show()

```

```
# Bar chart for categorical variable
df["Gender"].value_counts().plot(kind="bar")
plt.title("Count of Students by Gender")
plt.show()

# Boxplot grouped by category
df.boxplot(column="Marks", by="Gender")
plt.title("Marks by Gender")
plt.suptitle("")  # remove default title
plt.show()

```

Lab tasks

Task 1:

Generate a 10×10 random integer matrix (values between 0–100).

1. Find the row with the highest sum.
2. Normalize all values in the matrix using min-max scaling.
3. Replace the diagonal elements with the row means.
4. As an extension, compute the determinant of the transformed matrix.

Task 2:

Write a vectorized NumPy function that takes a 1D array as input. The function should:

1. Return the difference of each element from the array mean.
2. Mark elements greater than 1 standard deviation away as "outlier".
3. Extend the function to handle 2D arrays by applying the logic column-wise.
4. Visualize the results by plotting the array and highlighting outliers.

Task 3:
Simulate rolling two dice 100,000 times using NumPy.

1. Estimate the empirical probability distribution of sums (2–12).
2. Compare the results with the theoretical probability distribution.
3. Compute and visualize the cumulative probability distribution.
4. Extend the simulation to three dice and analyze how the distribution changes.

Task 4:
Given a dataset of students with columns [Name, Gender, Subject, Marks]:

1. Find the top scorer in each subject.
2. Compute the average marks per gender per subject.
3. Create a pivot table showing subjects vs. gender average marks.
4. Extend the task by finding the overall top 5 students across all subjects.

https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?utm_source=chatgpt.com

Task 5:
You are given a messy CSV file with missing values, mixed data types, and extra spaces. Your task is to:

1. Standardize and clean column names.
2. Convert numeric columns stored as text into proper numeric types.
3. Handle missing values using multiple strategies:
    - Fill numeric columns with the mean/median.
    - Fill categorical columns with the mode.
    - Drop rows with excessive missing values.
4. Extend the cleaning by removing duplicate rows and ensuring consistent formatting in categorical values.

Can take any random data
Task 6:
You are provided with a dataset containing sales data with columns [Date, Product, Sales]. Your task is to:

1. Convert the Date column into datetime format.
2. Calculate the monthly total sales.
3. Plot the sales trend over time and identify the month with the highest spike in sales.
4. Extend the analysis by identifying the top 3 products contributing to sales spikes and visualize their contribution.

https://www.kaggle.com/datasets/kyanyoga/sample-sales-data?utm_source=chatgpt.com
Task 7 :
A school has collected data on students, including their gender, study hours, attendance percentage, and marks in different subjects. The administration wants to know which factor has a stronger influence on academic performance: study hours or attendance. Your task is to explore the dataset, calculate group-wise averages, detect outliers in marks, and use visualizations to analyze how study hours and attendance relate to overall performance.
https://www.kaggle.com/datasets/larsen0966/student-performance-data-set?utm_source=chatgpt.com

Task 8:
A hospital maintains records of patients with details such as age, gender, blood pressure, cholesterol levels, and whether they are diagnosed with heart disease. The hospital management is interested in understanding the variation in cholesterol levels across age groups, identifying anomalies in blood pressure readings, and analyzing how cholesterol levels are related to the likelihood of heart disease. Perform an exploratory data analysis and present your findings with suitable charts and insights.
https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?utm_source=chatgpt.com

Task 9:
An e-commerce company provides transaction-level sales data that includes the order date, product, category, quantity sold, and sales amount. The management wants to identify sales trends over time, find the best- and worst-performing months, and discover which product categories generate the most revenue. Using EDA, explore the dataset and visualize patterns to help the company make better sales and marketing decisions.
https://www.kaggle.com/datasets/carrie1/ecommerce-data?utm_source=chatgpt.com

Task 10:
A tech firm has compiled data about its employees, including department, years of experience, projects completed, and performance scores. The HR team wants to compare employee productivity across departments, analyze whether experience is associated with productivity, and determine which department has the greatest variation in performance scores. Conduct an exploratory analysis to reveal these patterns and provide insights that can help improve employee management.
Random data from kaggle