Assignment 1 - Python for Data Science

MCQ Questions (1 mark each)

Q1. Which of the following Python libraries is most commonly used for data manipulation and analysis?
a) NumPy
b) Matplotlib
c) Pandas
d) Seaborn

Answer: c) Pandas

Q2. What is the output of the following Python code?

```
import numpy as np
arr = np.array([1, 2, 3, 4])
print(arr.mean())
```

a) 2
b) 2.5
c) 3
d) Error

Answer: b) 2.5

Descriptive Questions (5–7 marks each)

Q3. Explain the difference between NumPy arrays and Python lists. Give two advantages of using NumPy arrays in data science.

Answer:
- Python Lists are general-purpose containers that can hold elements of different data types, such as integers, floats, strings, etc. Operations on lists are relatively slower because they are implemented at a higher level.
- NumPy Arrays are specialized containers for numerical data, designed for fast computations and efficient memory usage. They allow element-wise operations and vectorized computations, making them ideal for data science tasks.

Advantages of NumPy arrays:
1. Performance: NumPy arrays are faster than Python lists because they use optimized C-based implementations and support vectorization.
2. Functionality: NumPy provides a wide range of mathematical and statistical functions

(mean, median, standard deviation, matrix operations, etc.) that are not directly available for lists.

Q4. What is data visualization in Python? Explain with an example using Matplotlib or Seaborn.

Answer:
Data visualization is the process of representing data graphically to identify patterns, trends, and insights easily. In Python, libraries like Matplotlib and Seaborn are commonly used for creating plots and charts.

Example using Matplotlib:

```
import matplotlib.pyplot as plt

x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

plt.plot(x, y, marker='o', color='blue')
plt.title('Simple Line Graph')
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.show()
```

This code generates a simple line graph showing a linear relationship between x and y. Visualization helps data scientists understand the data distribution and make better decisions.