Department of Data Science and Machine Learning

Half Term Examination

Course Title: Data Science and Machine Learning

Course Code: DS101

Duration: 3 Hours

Total Marks: 100

Instructions:

- **Attempt all questions.**
- **Read each question carefully before answering.**
- **All code should be properly commented and indented.**
- **Use appropriate data structures and algorithms where necessary.**
- **Marks are indicated next to each question.**
- **Assume any necessary imports (e.g., `import numpy as np`, `import pandas as pd`).**

Question 1: Python Fundamentals (15 Marks)

(a) Control Structures (7 Marks)

Write a Python function `fizz_buzz(n)` that prints numbers from 1 to `n`. For multiples of three, print "Fizz" instead of the number; for multiples of five, print "Buzz"; for numbers which are multiples of both three and five, print "FizzBuzz".

(b) Functions and Scope (8 Marks)

Explain the concept of variable scope in Python with an example. Include the differences between local, global, and nonlocal variables in your explanation.

Question 2: Data Structures in Python (20 Marks)

(a) Advanced List Manipulation (10 Marks)

Given a list of integers:

```
nums = [3, 6, 2, 7, 5, 6, 8, 5, 8, 3, 7]
```

Perform the following tasks:

- **Remove duplicates** from the list without using built-in functions that directly perform this operation. *(5 Marks)*
- **Sort** the list in ascending order. *(2 Marks)*
- **Slice** the sorted list to obtain a sublist containing the middle three elements. *(3 Marks)*

(b) Dictionary Comprehensions (10 Marks)

Using dictionary comprehension, create a dictionary that maps each character in the string `s = "Data Science"` to its corresponding ASCII value.

Question 3: NumPy Operations (15 Marks)

(a) Array Manipulations (10 Marks)

- Create a NumPy array of shape `(4, 4)` with values ranging from 1 to 16. *(2 Marks)*
- Reshape the array into an `(8, 2)` array. *(2 Marks)*
- Compute the **dot product** of the original `(4, 4)` array with its transpose. *(3 Marks)*
- Explain what the result represents. *(3 Marks)*

(b) Type Casting and Broadcasting (5 Marks)

Given a NumPy array:

```
a = np.array([1.5, 2.3, 3.7, 4.6])
```

- Convert it to an array of integers.
- Demonstrate how broadcasting works by adding this integer array to a 2D array:

```
b = np.array([[10], [20], [30], [40]])
```

Question 4: Data Analysis with Pandas (25 Marks)

You are provided with a CSV file `transactions.csv` containing the following columns: `'TransactionID'`, `'CustomerID'`, `'ProductID'`, `'Quantity'`, `'Price'`, `'TransactionDate'`.

(a) Data Loading and Cleaning (10 Marks)

- Load the dataset into a Pandas DataFrame. *(2 Marks)*
- Check for missing values and handle them appropriately (e.g., fill with mean, drop rows). Explain your choice. *(5 Marks)*
- Convert `'TransactionDate'` to datetime objects and extract the month into a new column `'Month'`. *(3 Marks)*

(b) Data Aggregation and Grouping (10 Marks)

- Calculate the **total revenue** (`Quantity` * `Price`) for each `'CustomerID'`. *(5 Marks)*
- Find the **top 5 customers** who have generated the most revenue. *(5 Marks)*

(c) Merging DataFrames (5 Marks)

Assume you have another DataFrame `customers.csv` with columns `'CustomerID'`, `'Name'`, `'Segment'`. Merge this DataFrame with the transactions DataFrame on `'CustomerID'` to analyze revenue by customer segment.

Question 5: Data Visualization (15 Marks)

Using the merged DataFrame from **Question 4**:

(a) Matplotlib Visualization (7 Marks)

Create a **line chart** showing the **monthly total revenue** over the period covered in the dataset. Include appropriate labels, title, and legend.

(b) Seaborn Visualization (8 Marks)

Use Seaborn to create a **histogram or density plot** of the distribution of transaction amounts. Comment on any skewness or anomalies in the data.

Question 6: Application of Data Analysis Concepts (10 Marks)

(a) Exploratory Data Analysis (EDA) (5 Marks)

Based on the `transactions.csv` dataset, identify any **two trends or patterns** that could be useful for a business. Provide supporting data or visualizations.

(b) Data Preprocessing (5 Marks)

Discuss the steps you would take to **prepare this dataset for a machine learning model**, such as linear regression to predict future sales. Mention at least **three preprocessing steps**.

Question 7: Capstone Coding Challenge (Bonus Question - Optional) (10 Marks)

Write a Python program that reads a text file `'book.txt'` and performs the following:

- Counts the **frequency of each word** in the text. *(4 Marks)*
- Identifies the **top 10 most frequent words** and displays them in a horizontal bar chart using Matplotlib or Seaborn. *(6 Marks)*

Total Marks: 100

Good luck!
