United International University
B.Sc. in Data Science
DS 3885: Data Wrangling

Mid Exam: Spring 2025 Time: 1 Hour 30 Minutes Marks: 20

Any examinee found adopting unfair means will be expelled from the trimester/program as per UIU disciplinary rules.

1. (a) You are assigned to collect product names and their corresponding prices from an e-commerce website for analysis. The HTML structure of the page is as follows:

```
<div class="product">
    <h2 class="product-name">Wireless Mouse</h2>
    <span class="price">$25.99</span>
</div>
<div class="product">
    <h2 class="product-name">Mechanical Keyboard</h2>
    <span class="price">$89.99</span>
</div>
```

Using Beautiful Soup in Python, write a code snippet to extract all product names and their prices. Describe the steps involved in the data collection process and explain how Beautiful Soup helps in navigating and parsing the HTML content. [5]

(b) Briefly explain the difference between `Selenium` and `Beautiful Soup`. When would you prefer one over the other in a web scraping task? [2]

2. You are tasked with cleaning and preparing a student performance dataset for analysis. The dataset contains details such as the student's name, grade, attendance, and course completion status. Below is a sample of the raw data stored in a CSV file:

```
Student ID, Name, Grade, Attendance, Completion Status
001, Alice, 85, 90, Completed
002, Bob, 75, NULL, Pend.
```

```
001, Alice, 85, 90, Complete
003, Charlie, 90, NULL, Completed
004, David, 78, 60, Complete
005, Eva, NULL, 80, Pending
```

(a) Identify the inconsistencies in this dataset and describe how you would resolve them. [2]

(b) Convert the "Completion Status" from textual data to numeric values using Label Encoding. [1]

(c) Fill in the missing Attendance value for Bob (ID 002) using the K-Nearest Neighbour Method (K=3) based on the available Attendance values of other students. [3]

3. You are given a small dataset of customer transactions from an e-commerce platform, where each transaction contains the total amount spent and the number of items purchased. The dataset consists of the following two features:

```
Amount Spent ($), Number of Items Purchased
200, 5
50, 2
2700, 6
20, 8
2500, 3
5500, 1
```

(a) Use K-means clustering to group the customers into two clusters based on the "Amount Spent" and "Number of Items Purchased". Perform the clustering for 2 iterations. [4]

(b) Check whether any outlier exists in the dataset or not. If yes, identify it and explain how it might affect the clustering result. [3]

$$\text{Outlier Threshold} = \mu + \sigma$$