DS5003 Data Engineering Lab

Midsemester exam, Date: Sep 27, 2025

Timing: 10:30 AM to 12:30 PM | Max Marks: 20

Instructions

1. Submit one .ipynb file containing all answers.
2. The name should be [student_name]_midsem.ipynb
3. Write the questions in separate text blocks before the answers.
4. Write justifications for your choices where needed.

---

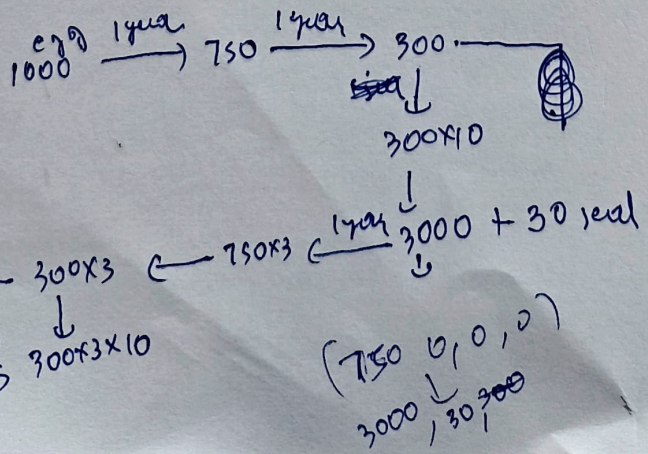
1. Given an array nums of size n, for each element find the next greater element (the first element to the right that is greater than the current element).If no such element exists, put -1.You have to do it in O(n) time complexity (3)

2. You are a biologist studying Atlantic Salmon. To suggest appropriate conservation strategies you have analyzed the lifecycle of the females of this species and have the following information: (3)

   - 75% of the eggs hatch into juveniles
   - 40% of the hatched eggs make it to the ocean and then back to their breeding grounds to lay eggs
   - 10% of the salmon who made it to the breeding adult stage end up surviving another year to lay eggs
   - Each breeding female lays on average 10 eggs in her first breeding season
   - If a female survives another year to breed again, she produces on average 5 eggs
   - No salmon survive their second year

   Assume you reintroduce 1000 Atlantic salmon eggs to an area from which they have gone extinct. Determine whether the total population will grow, shrink or stay the same over the years. Plot the change in the total population and the population age structure (proportion of each life stage in the population) over the years to form your conclusions.

   Hint: Let n_0: number of eggs, n_1: migrating salmon (juvenile to adulthood), n_2: breeding adult, n_3: surviving breeding adult. Then you can create a matrix L to calculate the population structure after a year recursively as $N_{t+1} = LN_t$, where $N_t = [n_0, n_1, n_2, n_3]$ at time t. The population structure after K years will be $N_K = L^K N_0$.



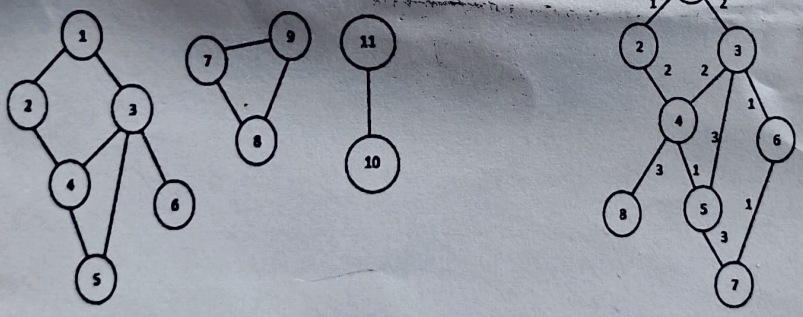
3. You are given datasets `Titanic.csv`, `Salary.csv`, and `Height.csv`. Using **Pandas**, **Matplotlib**, answer the following: (7)

- Load the `Salary.csv` and `Height.csv` datasets. Compute the 95% confidence interval for the mean of the `Salary` and `Height` columns using t distribution .(2 marks)
- Plot the Probability Density Function (PDF) for the `Salary` and `Height` columns. Detect and remove outliers using the z-score method if the PDF is approximately normal, otherwise use the IQR method. Re-plot the PDF for both columns after outlier removal.(2 marks)

  Note :For the Height column draw the PDF using either a line plot or a scatter plot.
- Load the `Titanic.csv` dataset. Plot a pie chart showing the proportion of passengers who survived vs. those who did not survive.(1.5 marks)
- Using the `Titanic.csv` dataset, apply **groupby** to find the average age of passengers by **class** and **sex**. Also, filter passengers who paid a fare above 100 and display their details. (1.5 marks)

4. (7)

- Implement DFS algorithm and demonstrate graph search on the graph in Figure 1a with source vertex as 1 and target vertex as 6. Print all the vertices in the order traversed until you found the target vertex. (**Constraint: Use a Python dictionary to store the adjacency list**)
- For the weighted graph given in Figure 1b, use Dijkstra's algorithm to find the diameter of the graph. (**Note**: the diameter of the graph is the longest path among the shortest paths between any two vertices.)



(a) (b)

Figure 1