Data Engineer Questions

This assignment has 10 questions, you have 1 week to complete as many questions as possible. You need to create a repository in github to submit an assignment. How well you construct the repository is a plus for you. Good luck to you!!!

SQL question: Using SQL language you feel comfortable with to solve.

Q.1: Assume that we have a table storing scores of athletes in a competition:

Performance*(AthleteId, Gender, Country, Score).*

Write an SQL to find the second highest score of athletes. (15 points)

Q.2: Assume that we have ***Customers****(id, name)* and ***Orders****(id, customerId)* tables.

Write an SQL query to report all customers who never order anything. (10 points)

Q.3: Assume that we have ***Employee****(id, name, salary, departmentId)* and ***Department****(id, name)*. A company's executives are interested in seeing who earns the most money in each of the company's departments. A **high earner** in a department is an employee who has a salary in the top three unique salaries for that department.

Write an SQL query to find the employees who are **high earners** in each of the departments. (20 points)

Coding question: You can use any language that you feel confident to solve the problem. Each question has many solutions, find the best solution and explain your solution, find the complexity of this solution (BigO)?

Q.4: We have an array of n elements A[1..n]. This array contains n different numbers from 0 to n. Given that there are totally n + 1 numbers from 0 to n, there is a missing number. (15 points)

Example:

- array: [0, 1, 2, 4]. The missing number is 3.
- array: [1, 4, 2, 3]. The missing number is 0.

Q.5: Given two sorted arrays `nums1` and `nums2` of size `m` and `n` respectively, return the median of the two sorted arrays. (20 points)

Example:

- nums1 = [1, 3, 6]; nums2 = [1, 2, 10] => Merged array: [1, 1, 2, 3, 6, 10]. Median is (2 + 3)/2 = 2,5.
- nums1 = [1, 3]; nums2 = [1, 2, 10] => Merged array: [1, 1, 2, 6, 10]. Median is 2.

Q.6: Given the head of a sorted linked list, *delete all duplicates such that each element appears only once*. Return *the linked list sorted as well*. (20 points)

Example:

- head = [2, 2, 5] => [2, 5]
- head = [1, 1, 3, 4, 4, 5, 6, 6] => [1, 3, 4, 5, 6]

Design question (Optional):

Q.7: We have a crawler that crawls websites in a list to find sensitive information (e.g., people talk or have opinions about our products.). Our list initially contains 100 websites called seeds. When a crawler visits a website, it can find several links to other websites. Depending on the information of the linking websites, they can be added in the list to revisit later (e.g., if they are related to the seeds or contain valuable information about what we want to know, i.e., sensitive information). It means that with the initial of 100 seeds, our list can be updated to include more websites. However, since our resources are limited, we want to maintain only up to 1000 websites. It means that in addition to 100 seeds, we can only maintain a maximum of 900 other websites. Design a solution to maintain this list of websites. Also justify your solution. (Max 10 points)

Hint: before the list grows to 1000 websites, you can freely add new websites to the list. However, when the list reaches 1000 websites, you need an algorithm to rank websites according to the usefulness of their information to our wanted information (i.e., sensitive information) and keep only those with high scores while removing those with low scores. E.g., when a new website is discovered and its score is higher than an existing one in the list, the new website will be replaced by the lower score in the list.

Mathematical statistics + other question (Bonus question):

Q.8: Assume that you are a rector of a university and you want to show to the public a statistics report for examinees in your university entrance exam, what is the best graph to use. Please justify your decision. (Max 2 points)

Q.9: We have three identical six-sided dice. We roll one dice first and the remaining two dice after that. What is the probability that the point obtained in the first roll is greater than the sum of the points obtained in the second roll. (2 points)

Q.10: Obtain the subway data set of NYC from this link:

https://github.com/andynganle/Data-for-Assignment

This data set records statistics of NYC subway riders in different weather conditions. Study and discuss any interesting features about the data set that you find. (Max 6 points)