Data Scientist Competency Assessment

2026

Instructions

This competency assessment consists of both a theory and practical component. You will be provided with a folder that contains:

1. An assessment document (45 marks).
2. Data files for the practical section.

Theory [14 marks]

- You may use the internet to assist you with your answers.
- It is not advised to use large language models to complete the assignment, and your grammar will not be used to judge your work.

Practical [31 marks]

- You may use the internet to assist you with your answers.
- It is not advised to use large language models to complete the assignment, and your grammar will not be used to judge your work.
- Please make use of any of the following programming languages: Python3, R, SQL, Java, Scala, etc.
- Please submit any software specific code or output used to complete the practical section.

Once you have completed the assessment, please email all the documents in your folder to the person who sent it to you within 4 hours, for example, if your assessment started at 9:00 am, you would be expected to hand-in your completed assessment by 13:00 pm on the same day.

Section A: Theory [14 marks]

1. An HbA1c level of 6.5% or more is considered high enough to diagnose diabetes, while values lower than 6.5% indicate healthy patients. A programmer incorrectly wrote the code to identify diabetics as: HbA1c > 6.5% and as such 300 patients with HbA1c = 6.5% were classified as healthy patients. In statistical terms, what could those 300 patients could be classified as?

   [1 mark]

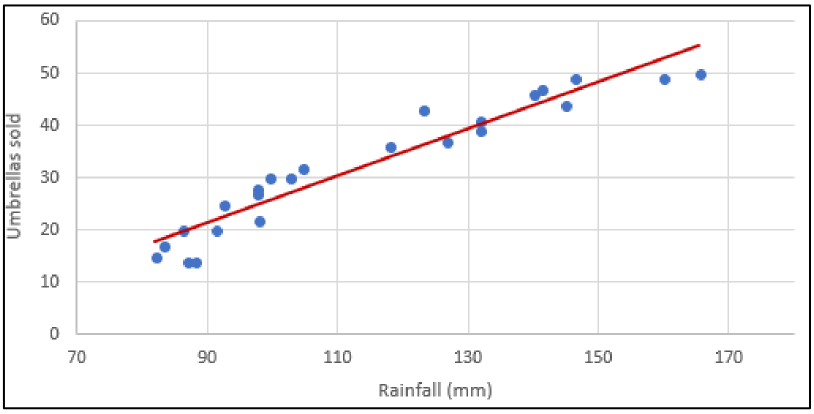
2. Knowing the medication a patient is taking often can indicate what health condition they have. For example, if a patient is regularly prescribed insulin, it is likely that they have chronic diabetes. However, in some cases patients that do not have chronic diabetes may be given insulin for short periods in hospital. Suppose there was a register of all patients who have been diagnosed with chronic diabetes by a doctor. You have been tasked with trying to reproduce this register based on insulin dispensing data alone. What statistical term would best describe the percentage of patients with chronic diabetes who were correctly identified as having the condition?

   [1 mark]

3. There are five people in an elevator and five floors in a building. What's the probability that each person gets off on a different floor? Show your thinking. Feel free to use paper and upload a photo of this if you'd prefer.
   [4 marks]

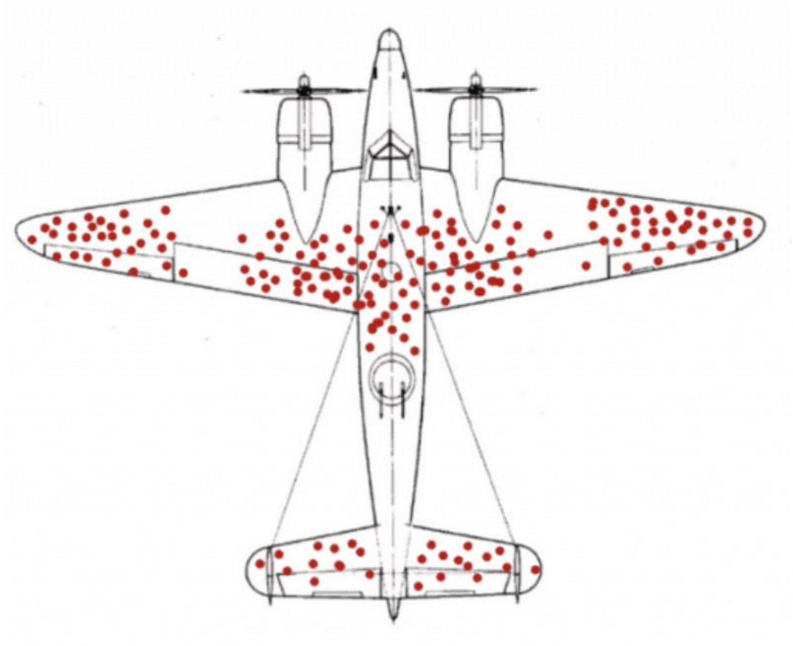
4. What assumptions would you need consider to be able to draw the red line below?

   [4 marks]



5. Consider the following problem and explain your reasoning:

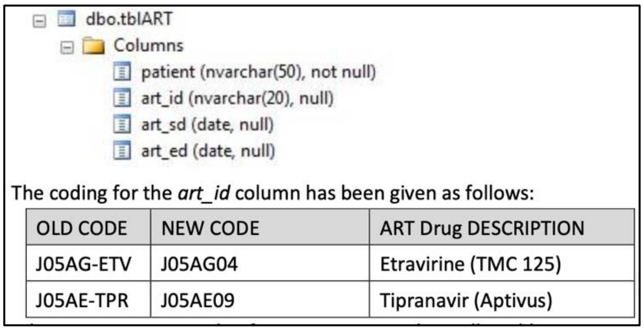
   Look at the picture below of the general pattern of holes from returning planes during the Korean War. What recommendation would you make to plane manufacturers in terms of where to add extra armour plating to ensure planes are able to withstand enemy gun fire without being shot down?



[4 marks]

Section B: Practical [26 marks]

1. There is a table called dbo.tblART in the SQL database, which contains antiretroviral therapy (ART) drug data. The codes for the ART drugs have changed and an update needs to be applied to the table. Please write and submit an example of a SQL statement that will enable you to update the information to reflect the new coding for the two drugs. If you do not know any SQL language you may code the update in the coding language of your choice. The ART table columns are as follows:



```
⊟ dbo.tblART
    ⊟ Columns
        patient (nvarchar(50), not null)
        art_id (nvarchar(20), null)
        art_sd (date, null)
        art_ed (date, null)
```

The coding for the *art_id* column has been given as follows:

OLD CODE	NEW CODE	ART Drug DESCRIPTION
J05AG-ETV	J05AG04	Etravirine (TMC 125)
J05AE-TPR	J05AE09	Tipranavir (Aptivus)

[2 marks]

2. You have been provided with data files:

   - *medicine_data.csv*:

   This contains medicine dispensing data for diabetic patients including patient ID, sex, date of birth, medicine name, and dispensing date.

   - *laboratory_data.txt:*

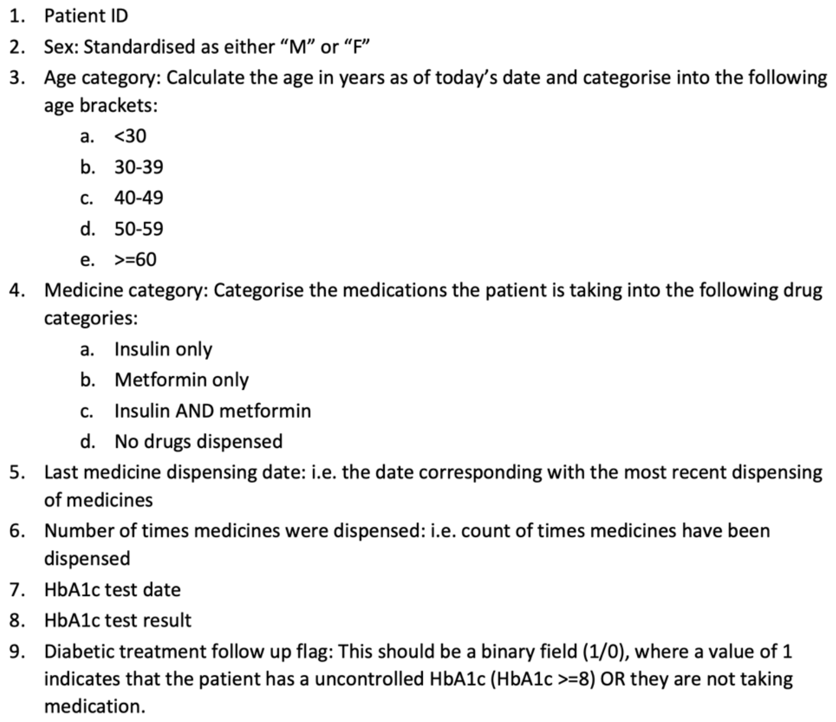
   This contains HbA1c test results for diabetic patients including patient ID, sex, date of birth, lab test type, lab test date, and lab test result.

   - *Note*:

   Please note that the patient ID is a unique identifier for the patient; i.e., each unique person has one "patient ID" that can be used to identify them.

   In this question you will be required to construct a wide dataset that could be used to track each individual patient's health. Your dataset should contain one line per patient and should contain the columns listed below. Construct your dataset using the language of your choice and submit both the completed dataset and code script.

   Dataset columns:



   1. Patient ID
   2. Sex: Standardised as either "M" or "F"
   3. Age category: Calculate the age in years as of today's date and categorise into the following age brackets:
      a. <30
      b. 30-39
      c. 40-49
      d. 50-59
      e. >=60
   4. Medicine category: Categorise the medications the patient is taking into the following drug categories:
      a. Insulin only
      b. Metformin only
      c. Insulin AND metformin
      d. No drugs dispensed
   5. Last medicine dispensing date: i.e. the date corresponding with the most recent dispensing of medicines
   6. Number of times medicines were dispensed: i.e. count of times medicines have been dispensed
   7. HbA1c test date
   8. HbA1c test result
   9. Diabetic treatment follow up flag: This should be a binary field (1/0), where a value of 1 indicates that the patient has a uncontrolled HbA1c (HbA1c >=8) OR they are not taking medication.

   [15 marks]

3. Using the above dataset, create a visualisation using the language of your choice that shows the distribution of male and female patients, using the age categories, who receive insulin only.

   [5 marks]

4. During your career, have you ever made a significant mistake in your analysis?
   a. How did you handle it and

    b. what did you learn from it?

[3 marks]

5. Describe a project where you had to work with a difficult team member.
    a. How did you handle the situation?

[3 marks]

6. Can you share an example of a time when you had to work under a tight deadline?
    a. How did you manage your tasks and deliver on time?

[3 marks]