# Data scientist test

## About The Health Economics Unit (HEU)

The HEU has been created to deliver impactful insights that power the future development of health and care delivery. Our experts support health system and industry leaders to make the best possible decisions and design the highest quality, most efficient and innovative services to improve the health of communities across the country.

The HEU is a specialist team hosted by the NHS Midlands and Lancashire Commissioning Support Unit (MLCSU), one of five CSUs across the UK responsible for providing a range of services and support to clients including clinical commissioning groups (CCGs), hospital trusts, local authorities, integrated care systems (ICSs) and other public bodies.

## SQL and database design test

The HEU has been asked to design and build a database to store patient information with the following requirements:

* Appointments: We want to be able to book appointments into the future, which are then updated with outcomes. The intention and outcomes of the appointments could vary, as could the member of staff carrying them out.
* Periods of care in a hospital: We want to be able to record when someone enters a care facility for a period of time, typically at least overnight using a bed.
* Identifiable information: We want to be able to record information that allows us to identify an individual & their characteristics

Please provide:

1. Any assumptions you made.
2. A description of how you mapped your design into a set of tables.
3. A relational schema.
4. A SQL script to create the tables.
5. A SQL script that would produce some data quality checking.

## Coding test

Using a programming language of your choice and the patient\_health.csv dataset provided, please perform the tasks below.

**Data Cleaning:**  
Identify and handle any missing or inconsistent data (e.g., missing values in Age, invalid dates).

* Ensure that the AppointmentDate field is in a consistent format and can be used for date calculations.
* Standardise categorical fields like Outcome (e.g., ensure consistent spelling/capitalisation).

**Descriptive Statistics and Insights:**  
Calculate:

* The no-show rate for each GP practice.
* The distribution of patients’ ages by diagnosis.
* The proportion of male and female patients for each diagnosis.

**Advanced Analysis:**  
Write code to prepare the dataset for a predictive model that would estimate the likelihood of a patient missing their appointment. The prepared dataset should include:

* Relevant features.
* A target variable.

Create at least one plot (e.g., bar chart, histogram) to visualise:

* The no-show rates by diagnosis.

A second visualisation of your choice to highlight another insight.

Please provide:

1. A script or notebook with your complete code.
2. Comments in your code explaining key steps.
3. A brief summary of your findings.

**Analysis plan development**

The Health Economics Unit (HEU) is working to understand and address the high rates of emergency leukaemia diagnosis, which have remained consistent over the past five years. This task is designed to evaluate your ability to create an actionable analysis plan to support early detection and intervention strategies for leukaemia, with the goal of improving patient outcomes and optimising healthcare system efficiency.

You are tasked with creating an analysis plan for a project aiming to address the challenge of emergency leukaemia diagnoses. The plan should be structured, logical, and actionable, reflecting a clear understanding of the data and analytical methods required.

The HEU has been provided with patient-level data from 2009–2024. The goal is to analyse patterns of emergency leukaemia diagnoses, identify opportunities for earlier detection, and evaluate the relationship between late diagnosis, treatment access, and outcomes.

**Your response must fit on one page of a Word document.** Use bullet points and concise phrasing where possible.

Problem Definition

* Summarise the primary objective of the project.
* List 1–2 key hypotheses.

Data and Assumptions

* Identify the key data needed.
* State any key assumptions.

Proposed Analyses

* Trend Analysis: Briefly describe how you would examine changes in leukaemia diagnosis rates over time.
* Patient Journey Analysis: Outline how you would explore the timeline from symptom onset to diagnosis and treatment.
* Subgroup Analysis: Note any key groups for comparison.

Challenges and Mitigation

* Identify one potential challenge and propose a simple solution.