Background

Business Domain

In the complex and critical landscape of UK healthcare, the National Health Service (NHS) provides a wide range of services to patients across the nation. These services span all aspects of healthcare, from routine checkups and medical tests to complex surgeries and ongoing medical management. The NHS collects and maintains a vast amount of data related to patients, healthcare professionals, facilities, treatments, and financial transactions. This data, if properly analysed, can be a goldmine of insights for improving healthcare delivery, resource allocation, and patient outcomes. However, [assume that] much of this data currently resides in disparate, unnormalised CSV files, making it difficult to gain a comprehensive view.

Imagine you are a data scientist working for the NHS. Recognising the crucial role of data in enhancing the quality and efficiency of healthcare, the Chief Executive Officer of the NHS Trust has established a Data Intelligence team, and appointed you as its lead data scientist.

Aim: The Data Intelligence team is tasked with undertaking a data science project leveraging the available healthcare data. The project's findings will be presented to key stakeholders within the NHS Trust, including hospital administrators, medical professionals, and policymakers. These findings will serve as a foundation for improving patient care, optimising resource utilisation, and driving data-informed decision-making across the organisation. Potential applications include predicting hospital readmissions, personalising treatment plans, identifying high-risk patients, and analysing the effectiveness of different medical interventions. Ultimately, the project aims to contribute to a more efficient, effective, and patient-centred healthcare system.

UK Healthcare Data Overview

The NHS maintains extensive records related to patient care, healthcare professionals, facilities, and financial transactions. These records, while currently stored in separate CSV files, represent a rich source of information for improving healthcare delivery and management. The data encompasses several key areas:

Patients and Demographics: The NHS collects demographic information about patients, including a unique patient identifier, name, date of birth, gender, address, and preferred healthcare providers (a preferred hospital, a preferred pharmacy, and a chosen insurance provider). This data is crucial for patient identification, tracking, and personalised care. Furthermore, the NHS maintains records of emergency contacts for each patient, including their relationship to the patient, contact identifier, name, phone number, and address. A patient can have one or more persons as emergency contacts, and a person can be emergency contact of one or more patients.

Healthcare Facilities and Professionals: Information about hospitals is maintained, including a unique hospital identifier, name, location, and contact number. A hospital consists of one or

more departments. The data about departments within each hospital includes a unique department identifier, the hospital they belong to, their name, and the head of the department. The department identifier is unique across all hospitals, not just within its own hospital. Healthcare professionals (professionals) work in a department within a hospital. They have a unique professional identifier, name, and their role (*nurse, midwife, surgeon,* etc.). This information is linked to the departments where they work. A professional can work in more than one department, all within the same hospital or in different hospitals.

Pharmacies provide the medications prescribed by professionals to patients. Data about pharmacies include a unique pharmacy identifier, name, location, contact number, operating hours, manager name, website, and services offered. Lastly, medical laboratories (labs) conduct medical tests (tests) for patients when recommended by a professional. Data about labs include a unique lab identifier, name, location, contact number, and the type of test they perform. Each lab is capable of conducting a single type of test.

Medical Encounters and Records: A patient can have an appointment with professionals for consultation at a department or undergo a surgery by a professional at a hospital. Information about patient appointments includes a unique appointment identifier, the patient, the professional, the date and time of the appointment, the department, and the appointment status (*completed, scheduled,* or *cancelled*). Information about surgeries performed on patients includes a unique surgery identifier, the patient, the professional who performed the surgery, the hospital where the surgery took place, the date, type of surgery, notes, post-operative care instructions, and the outcome. Patients can also take tests upon recommendation by healthcare professionals at a lab. The data about tests include a unique test identifier, the patient, the professional who recommended the test, the test name, results, date, the lab where the test is performed, and the billing type (whether *NHS* or *private*).

The NHS maintains detailed medical records for each patient, identified by a unique record identifier. A new record is created for a patient for each *completed* appointment, surgery, and test. These records contain information about diagnoses, notes from professionals, and links to specific appointments, surgeries, and tests.

Medications and Prescriptions: Patients may be prescribed medications by professionals during an appointment. Data about medications include a unique medication identifier, name, manufacturer, dosage form (Tablet, Capsule, Syrup, etc.), strength (250mg, 500mg, 5mg/5ml, 2%, etc.), and price. Prescriptions issued to patients after completing an appointment have a unique prescription identifier and the recorded details include the patient, the professional who prescribed the medication, the associated medical record, and the pharmacy where the prescription is served. A prescription can have one or more medications for the patient. Details about each medication within a prescription include the recommended dosage of the medication, quantity, total billing amount of the particular medication, start date, and end date. The total billing amount of medications within a prescription may be covered by insurance claims individually if the patient has a medical insurance provider.

Billing and Insurance: The NHS maintains records of medical billings and insurance claims,

including billings for medications and services. Medication billings are maintained with prescriptions (discussed earlier). Service billings cover appointments and surgeries at a private hospital, and tests of *private* billing type. Bills of all other types of services are directly covered by NHS. The details include a unique billing identifier, the patient, the related appointment, surgery, or test, the amount, payment status (*paid, pending, partially paid*), amount paid, and payment date.

There exist insurance providers who offer medical insurance to patients for their bills. Information about insurance providers includes a unique insurance provider identifier, name, contact number, and coverage type (a list of one or more types such as *dental*, *vision*, etc.). For simplicity, treat the coverage type as a single-valued attribute (a list) even when an insurance provider provides a combination of multiple types. For patients who have got medical insurance from an insurance provider, insurance claims can be raised to pay their medication and service bills. Data about insurance claims include a unique claim identifier, the associated prescription or service billing, the insurance provider, claim status, claim status reason, amount claimed, approved amount, date, and approval date.

This data, when properly integrated and analysed, can provide valuable insights into various aspects of healthcare, enabling the NHS to improve patient care, optimise resource allocation, and enhance the overall efficiency of the healthcare system.

Tasks

You are required to design and build solutions based on the UK healthcare scenario discussed (pages 3-5) and the data provided (pages 10-16). There are four specific tasks: Task 1 – Task 4. In particular,

- Task 1: conduct domain analysis;
- Task 2: design a database for the provided scenario and data;
- Task 3: design and execute a data analysis plan to derive actionable insights from the provided healthcare data, addressing the business domain and needs.
- Task 4: answer the provided ethical and analytical questions.

A detailed description of each task is provided below.

Your findings and code should be presented as a Jupyter notebook and include the tasks described below. The dataset (present in six files: *Appointments_Data.csv, Prescription_Billing_Insurance_Data.csv, Service_Billing_Insurance_Data.csv, Medical_Appointments_Data.csv, Medical_Surgeries_Data.csv and Medical_Tests_Data.csv*) and a template of the Jupyter notebook with the four tasks are available on VLE (*VLE → DATA → Assessment → Assessment 2024-25 resources*). Further formatting details and word limit per task are given below.

Task 1: Domain Analysis (5 Marks)

Given the business domain and the data overview presented above, provide a concise description of

- the business problem and its significance to the relevant sector;
- the link between the business problem and the field of data science;
- the main areas of investigation;
- potential ideas and solutions.

Word Limit: Your answer to this question must not exceed **300 words**.

Task 2: Database Design (30 Marks)

(10 marks) Design a conceptual database schema for the given NHS context, represented as an entity-relationship (ER) diagram using Chen's notation. Your ER diagram should capture all the essential entities, attributes, primary keys, relationships, and cardinalities, necessary to model the healthcare operations described in the scenario .

The healthcare data currently exists in the form of six csv files called *Appointments_Data.csv, Prescription_Billing_Insurance_Data.csv, Service_Billing_Insurance_Data.csv, Medical_Appointments_Data.csv, Medical_Surgeries_Data.csv and Medical_Tests_Data.csv*, provided on VLE (*path given in page 5*). These files have all the existing records. The tables available in the csv files are unnormalised. The information about the different columns in them is given in Tables 1, 2, 3, 4, 5 and 6, respectively.

(10 marks) Normalise the provided tables to the Third Normal Form (3NF), minimising data redundancy and ensuring data integrity. Demonstrate the steps involved in achieving 3NF, showing how you decomposed the tables through 1NF and 2NF.

(10 marks) Finally, implement your 3NF schema in an SQLite database using SQL. Your answer should include the SQL statements needed to accomplish this step and populate the final tables with the appropriate data.

Your submission should include the final SQLite database file.

Your answer should clearly cover the following:

- Any assumptions you are making about the given scenario;
- The designated primary and foreign keys, existing relationships, and identified functional dependencies;
- The steps followed and justifications for the decisions made.

World Limit: Your answer to this question must not exceed **500 words**. This limit applies only to the explanations. There is no limit on any associated code, SQL statements, figures and tables.

Task 3: Research Design, Implementation, and Results (45 Marks)

Using the database schema designed in Task 2, develop, implement, and analyse three distinct modelling solutions **(15 marks each)** to address the Data Intelligence team's aim (as described in the scenario). The three solutions must collectively cover all three of the following categories: inferential statistics, supervised learning, and unsupervised learning, with each solution primarily focusing on one or a combination of these categories. The solutions should be of sufficient

complexity to demonstrate a comprehensive understanding of the data and the problem. For each solution, include:

- **Problem**: Clearly and concisely state the specific problem within the NHS context that your solution addresses.
- **Solution**: Detail the design of your solution, including the specific techniques used and how they are combined. Ensure that your design incorporates information from multiple tables in the database where relevant.
- **Justification**: Explain why the selected inferential statistics, supervised learning algorithms, and/or unsupervised learning algorithms are appropriate for the specific problem being addressed.
- **Implementation**: Provide well-commented and organised code (including SQL queries) used to implement your solution. Clearly indicate and justify any modifications made to the database schema or data. Ensure that your code is reproducible.
- **Results**: Present your findings in a clear and concise manner, using appropriate visualisations (charts, graphs, tables) as appropriate. Critically analyse your results, discussing how they can help the NHS address the stated problem.
- **Limitations**: Discuss any limitations of your solution, including potential biases in the data, assumptions made, or areas where the solution could be improved.

World Limit: Your answer to this question must not exceed **400 words** per solution. This limit applies only to the explanations. There is no limit on any associated code or figures.

Task 4: Ethics and Analysis (10 Marks)

I. **(5 marks)** Discuss the ethical implications of your modelling solutions given in Task 3. How can these ethical challenges be mitigated in a real-world NHS setting? Your answer to this question must not exceed **200 words**.

II. **(5 marks)** Write a Python script using SQL to analyse the database from Task 2 and generate results showing: (a) The distribution of hospitals across cities. (b) For each hospital, its name, city, the number of departments, and the number of patients who prefer that hospital, using outer join. Sort the hospitals within each city by the number of preferred patients in descending order.

Overall Academic Quality (10 Marks)

10 marks are allocated for the clarity, cohesiveness and reproducibility of your answers (both text and code) across all tasks with appropriate, relevant and effective analysis and presentation of the results.

Deliverables

You should submit the following to the submission point on VLE Ultra:

1. the SQLite database produced in Task 2;
2. the completed Jupyter notebook (both .ipynb and HTML files) containing solutions for all the tasks. A template has been provided on VLE;
3. any figures or diagrams that are included in your answers in the Jupyter notebook.

Table 1: Description of columns in *Appointments_Data*

Attribute	Description	Remarks
Appointment_ID	Unique identifier for the appointment.	
Patient_ID	Unique identifier for the patient.	
Appointment_Date	Date when the appointment occurred.	YYYY-MM-DD.
Appointment_Time	Time when the appointment occurred.	
Appointment_Status	Status of the appointment	*Completed, Scheduled,* or *Cancelled*
Patient_Name	Name of the patient.	
Patient_Date_Of_Birth	Patient's date of birth.	YYYY-MM-DD.
Patient_Gender	Patient's gender.	M: Male, F: Female.
Patient_Address	Address of the patient.	
Patient_Preferred_Hospital_ID	Preferred hospital ID of the patient.	
Patient_Preferred_Pharmacy_ID	Preferred pharmacy ID of the patient.	
Patient_Preferred_Insurance_Provider	Preferred insurance provider of the patient.	
Emergency_Contact_Relationship	Relationship to the patient of the emergency contact.	
Emergency_Contact_Name Emergency_Contact_Phone	Name of the emergency contact. Phone number of the emergency contact.	
Emergency_Contact_Address	Address of the emergency contact.	
Professional_Name	Name of the professional handling the appointment.	
Professional_Role	Role of the professional	e.g., *Neurologist, Speech Therapist*.
Department_ID	Unique identifier for the department in which the professional works.	
Department_Name	Name of the department.	Based on the speciality.
Head_of_Department	Name of the head of the department.	
Hospital_ID	Unique identifier for the hospital.	

Attribute	Description	Remarks
Hospital_Name	Name of the hospital.	Contains the name of the city in which it is located and the type of the hospital.
Hospital_Location	Location of the hospital.	
Hospital_Contact	Phone number of the hospital.	

Table 2: Description of columns in *Prescription_Billing_Insurance_Data*

Attribute	Description	Remarks
Prescription_ID	Unique identifier for the prescription.	
Record_ID	Identifier for the medical record associated with the prescription.	
Pharmacy_ID	Unique identifier for the pharmacy.	
Prescription_Detail_ID	Unique identifier for the prescription detail.	
Medication_ID	Unique identifier for the medication prescribed.	
Medication_Dosage	Dosage of the medication prescribed.	
Medication_Quantity	Quantity of the medication prescribed.	
Total_Medication_Billing_Amount	Total billing amount for the prescribed medication.	
Dosage_Start_Date	Start date for the prescribed medication dosage.	
Dosage_End_Date	End date for the prescribed medication dosage.	
Pharmacy_Name	Name of the pharmacy.	
Pharmacy_Location	Location of the pharmacy.	
Pharmacy_Contact	Contact number of the pharmacy.	
Pharmacy_Email	Email address of the pharmacy.	
Pharmacy_Operating_Hours	Operating hours of the pharmacy.	
Pharmacy_Manager_Name	Name of the pharmacy manager.	
Pharmacy_Website	Website URL of the pharmacy.	

Attribute	Description	Remarks
Pharmacy_Services_Offered	Services offered by the pharmacy.	It is actually multi-valued, but for simplicity treat it as a single-valued list of multiple services.
Medication_Name	Name of the medication.	
Manufacturer	Manufacturer of the medication.	
Medication_Dosage_Form	Form of the medication.	e.g., *Tablet, Syrup.*
Medication_Strength	Strength of the medication	e.g., *500mg.*
Medication_Price	Price of the medication.	
Claim_ID	Unique identifier for the insurance claim.	
Claim_Status	Status of the insurance claim.	
Claim_Status_Reason	Reason for the claim status.	
Claim_Amount	Total amount of the claim.	
Approved_Amount	Amount approved by the insurance provider.	
Claim_Date	Date of the insurance claim.	
Approval_Date	Date when the insurance claim was approved.	
Insurance_Provider_ID	Unique identifier for the insurance provider.	
Insurance_Provider_Name	Name of the insurance provider.	
Insurance_Provider_Contact	Contact number of the insurance provider.	
Insurance_Provider_Coverage_Type	Types of coverage offered by the insurance provider.	It is actually multi-valued, but for simplicity treat it as a single-valued list of multiple coverage types.

Table 3: Description of columns in *Service_Billing_Insurance_Data*

Attribute	Description	Remarks
Claim_ID	Unique identifier for the insurance claim.	
Claim_Status	Status of the insurance claim.	
Claim_Status_Reason	Reason for the claim status.	
Claim_Amount	Total amount of the claim.	

Attribute	Description	Remarks
Approved_Amount	Amount approved by the insurance provider.	
Claim_Date	Date of the insurance claim.	
Approval_Date	Date when the insurance claim was approved.	
Insurance_Provider_ID	Unique identifier for the insurance provider.	
Insurance_Provider_Name	Name of the insurance provider.	
Insurance_Provider_Contact	Contact number of the insurance provider.	
Insurance_Provider_Coverage_Type	Type of coverage offered by the insurance provider.	It is actually multi-valued, but for simplicity treat it as a single-valued list of multiple coverage types.
Service_Billing_ID	Unique identifier for the service billing.	
Appointment_ID	ID of the appointment linked to the service billing.	
Surgery_ID	ID of the surgery linked to the service billing.	Surgeries carried out at a private hospital are billed to the patient, which can be covered by insurance or directly paid.
Test_ID	ID of the test linked to the service billing.	Only Private tests are billed, NHS tests are automatically covered by NHS.
Service_Billing_Amount	Amount billed for the service.	
Service_Billing_Payment_Status	Payment status of the service billing.	*Paid, Pending,* or *Partially Paid*
Service_Billing_Amount_Paid	Amount paid for the service.	
Service_Billing_Payment_Date	Date when payment for the service was made.	

Table 4: Description of columns in *Medical_Appointments_Data*

Attribute	Description	Remarks
Record_ID	Unique identifier for the medical record.	
Diagnosis	Medical diagnosis related to the test, surgery, or appointment.	
Notes	Notes related to the medical record.	
Appointment_ID	Identifier for the appointment linked to the medical record.	NULL, if the medical record is related to a surgery or a medical test.
Surgery_ID	Identifier for the surgery linked to the medical record.	NULL, if the medical record is related to an appointment or a medical test.
Test_ID	Identifier for the medical test linked to the record.	NULL, if the medical record is related to an appointment or a surgery.
Patient_ID	Unique identifier for the patient.	
Appointment_Date	Date when the appointment occurred.	
Appointment_Time	Time when the appointment occurred.	
Appointment_Status	Status of the appointment.	*Completed, Scheduled,* or *Cancelled*

Table 5: Description of columns in *Medical_Surgeries_Data*

Attribute	Description	Remarks
Record_ID	Unique identifier for the medical record.	
Diagnosis	Medical diagnosis related to the test, surgery, or appointment.	
Notes	Notes related to the medical record.	
Appointment_ID	Identifier for the appointment linked to the medical record.	
Surgery_ID	Identifier for the surgery linked to the medical record.	

Attribute	Description	Remarks
Test_ID	Identifier for the medical test linked to the record.	
Patient_ID	Unique identifier for the patient.	
Surgery_Professional_ID	Professional ID for the surgery.	
Surgery_Hospital_ID	Hospital ID where the surgery was performed.	
Surgery_Date	Date when the surgery was conducted.	
Surgery_Type	Type of surgery performed.	
Surgery_Notes	Notes related to the surgery.	
Surgery_Post_Operative_Care	Details of post-operative care for the surgery.	
Surgery_Outcome	Outcome of the surgery.	

Table 6: Description of columns in *Medical_Tests_Data*

Attribute	Description	Remarks
Record_ID	Unique identifier for the medical record.	
Diagnosis	Medical diagnosis related to the test, surgery, or appointment.	
Notes	Notes related to the medical record.	
Appointment_ID	Identifier for the appointment linked to the medical record.	
Surgery_ID	Identifier for the surgery linked to the medical record.	
Test_ID	Identifier for the medical test linked to the record.	
Patient_ID	Unique identifier for the patient.	
Test_Recommended_By_Professional_ID	ID of the professional who recommended the test.	
Test_Name	Name of the medical test.	
Test_Results	Results of the medical test.	
Test_Date	Date when the test was conducted.	
Lab_ID	Unique identifier for the lab where the test was conducted.	
Test_Billing_Type	Billing type for the test.	*NHS or Private*.
Lab_Name	Name of the laboratory.	

Attribute	Description	Remarks
Lab_Location	Location of the laboratory.	
Lab_Contact	Contact details for the lab.	
Lab_Type	Type of laboratory.	e.g., *Forensic Science, Cardiology*.