PRIFYSGOL ABERYSTWYTH UNIVERSITY

SEMESTER TWO 2026

COMPUTER SCIENCE

CSM6720 : ADVANCED DATA ANALYTICS

Time allowed: **2 HOURS**

Answer all **THREE** questions

All questions carry equal marks.

YOU MAY BRING THE FOLLOWING INTO THE EXAMINATION:
UP TO 3 SHEETS OF NOTES ON A4 PAPER (WHICH MAY BE DOUBLE-SIDED). THERE ARE NO ADDITIONAL RESTRICTIONS ON CONTENT OR FORMATTING OF THESE NOTES.
NO ELECTRONIC DEVICES.

CALCULATORS ARE PERMITTED (CASIO FX991, FX83 OR FX85 MODELS ONLY)

STUDENTS MAY SUBMIT ANSWERS TO THIS PAPER IN EITHER WELSH OR ENGLISH

Once seated, students may complete the front cover of the answer book(s) and the attendance slip ONLY.

DO NOT TURN THE EXAMINATION QUESTION PAPER OVER UNTIL INSTRUCTED TO DO SO.

1. **This question is related to relational and non-relational databases.**

Hywel Dda University Health Board is developing a new integrated digital platform that will support three main services.

a) The first service is a **Patient records management system.** In this system, each patient has structured demographic information (name, date of birth, NHS number, address). Each patient may have multiple appointments. Each appointment is linked to a clinician and includes diagnosis codes and prescriptions. The system must enforce strong data consistency. The system must support complex queries such as: *"List all patients treated by Dr Evans in the past 6 months with a diagnosis of hypertension"*.'

(i) For this system, would you use relational or non-relation database? Justify your choice. [3 marks]

(ii) Align your choice of database type with the intended system. Include a discussion on the data structure, scalability and performance, consistency guarantees, query complexity and the flexibility of schema of your chosen technology. [6 marks]

(iii) Give an example and explanation of ONE risk associated with your choice. [2 marks]

b) The second service is a **Remote health monitoring system.** This system will enable patients to use wearable devices that stream heart rate, movement, and sleep data every second. The system receives millions of records per day, so the architecture must support high write throughput. The data structure may evolve over time as new sensors are added. The architecture should allow multiple machines processing the data in parallel so that performance remains stable as the number of patients increases.

Marking criteria: 1 mark for correctly identified database type and 2 marks for justification

(i) For this system, would you use relational or non-relation database? Justify your choice. [3 marks]

(ii) Align your choice of database type with the intended system. Include a

discussion on the data structure, scalability and performance, consistency guarantees, query complexity and the flexibility of schema of your chosen technology. [6 marks]

(iii) Give an example and explanation of ONE risk associated with your choice. [2 marks]

c) The third system is a **Patient feedback and messaging platform.** This will allow patients to send messages to clinicians. Messages may include text, images, voice recordings and file attachments. Conversation format may change between health departments. The structure of stored data may change over time.

(i) For this system, would you use relational or non-relation database? Justify your choice. [3 marks]

(ii) Align your choice of database type with the intended system. Include a discussion on the data structure, scalability and performance, consistency guarantees, query complexity and the flexibility of schema of your chosen technology. [6 marks]

(iii) Give an example and explanation of ONE risk associated with your choice. [2 marks]

2. **This question is related to entity relationship modelling and data analysis.**

A technology company is developing a large-scale Smart Home Analytics Platform. The system collects data from thousands of connected homes. Each home contains multiple smart devices and sensors:

- Each Home has HomeID, Location and HouseholdSize.
- Each Device has DeviceID, DeviceType (e.g., thermostat, light, washing machine, smart plug), InstallationDate and PowerRating.
- Each Sensor has SensorID SensorType (e.g., temperature, motion, humidity, energy meter) and Unit.
- Devices belong to one home. Sensors are attached to one device. Sensors generate time-stamped readings (DateTime, Value). Homes may define multiple Automation Rules (e.g., *"If motion detected, then turn light on"*). Each rule may involve multiple devices.

The company wants to analyse:

- Hourly and daily energy usage patterns
- Abnormal energy consumption
- Device interaction chains
- Effectiveness of automation rules
- Comparison of energy efficiency across locations

a) Design an Entity Relationship Diagram (ERD) for the system described above. Your answer must include all entities, all attributes, all relationships and correct cardinalities. [12 marks]

b) Using your ERD design, answer the following:

(i) Explain how your model enables analysis of hourly and daily energy consumption patterns per home. [3 marks]

(ii) Describe how you would identify homes experiencing unusually high energy usage compared to similar homes. [3 marks]

(iii) Explain how you would evaluate whether automation rules reduce overall energy consumption. [3 marks]

(iv) Explain how your model supports comparison of energy efficiency across different geographic locations. [3 marks]

c) The analytics system generated the following summary showing average daily energy consumption (kWh) for two groups of homes: *Homes without automation rules* and *Homes with automation rules enabled*. The figure below shows the average daily consumption over a 7-day period. Based on the figure, what conclusion can be drawn regarding the impact of automation rules? Suggest ONE possible explanation for the observed pattern. Explain how your ERD supports generating this comparison. [5 marks]

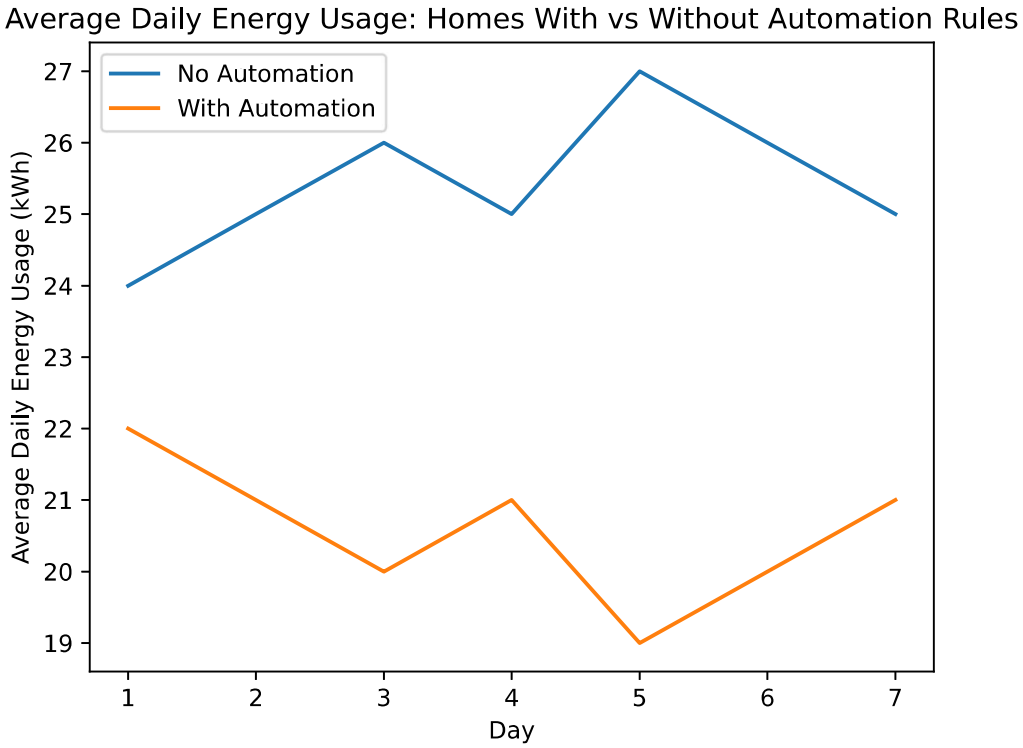


Figure 1: Comparison of average daily energy consumption (kWh) over a seven-day period for homes with and without automation rules enabled.

d) Below is a simplified graph showing average hourly energy usage (kWh) for three smart homes over 24 hours. Identify which home shows the most abnormal

behaviour and justify your answer. Suggest ONE plausible smart home explanation for the observed pattern. [4 marks]

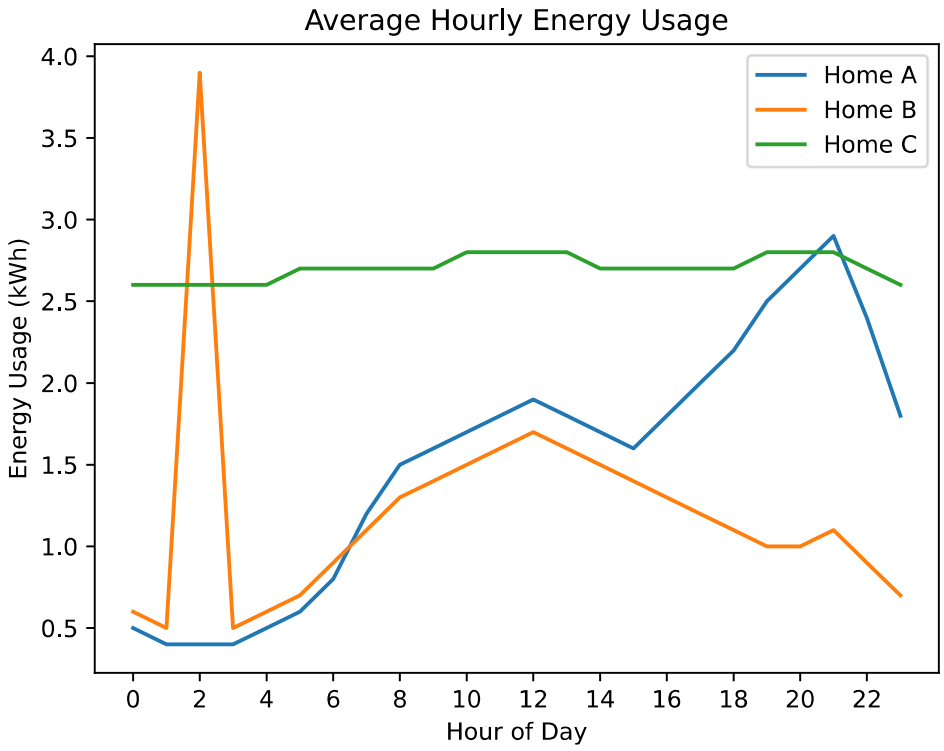


Figure 2: Average Hourly Energy Usage (kWh) for Three Smart Homes

3. **This question is related to evaluation of NoSQL data models.**

A global social networking platform is being designed. The system must support:

- User profiles (name, age, country, interests)
- Posts (text, images, timestamps)
- Comments on posts
- “Likes” on posts and comments
- Friend relationships between users
- The ability to recommend new friends based on mutual connections
- The ability to display a user’s feed quickly
- The ability to analyse trending topics across millions of posts

The system is expected to handle very high read and write volumes and operate at global scale.

a) For each of the following NoSQL data models:

- key-value;
- document-oriented; and
- graph,

identify and explain:

- the core structural concept of the model;
- how data is stored and accessed; and
- ONE typical use case where it performs well.

[9 marks]

b) For each of the three NoSQL data models, evaluate how suitable it would be for the social networking platform described above. Provide at least TWO strengths and ONE limitation for each model in this context. Your answer must refer directly to the system requirements. [12 marks]

c) The management at the social media company decide to prototype two data models. Upon testing, the following is discovered:

- **Prototype System 1.** The system handles user posts efficiently, but generating friend recommendations based on mutual connections has become increasingly slow as the network grows. Queries that involve multiple levels of user relationships require complex joins and heavy computation in the application layer.
- **Prototype System 2.** The database efficiently handles relationship queries, but storing and retrieving large volumes of posts and comments is becoming inefficient. Feed generation queries require traversing many small connected nodes, leading to high latency.

(i) What data model have they likely chosen for *Prototype System 1*? Explain why these problems indicate that data model, and suggest a better model for that part of the system. [6 marks]

(ii) What data model have they likely chosen for *Prototype System 2*? Explain why these problems indicate that data model, and suggest a better model for that part of the system. [6 marks]

END OF PAPER