Daffodil International University
Faculty of Science & Information Technology
Department of Software Engineering
Final Examination, Summer 2026
Course Code: DS331; Course Title: Introduction to Data Science, Data Management & Analysis
Sections & Teachers: 41 & (A,B,C) MSA, (D,E) IAT, (F,G) SSI

Time: 2 Hour **Marks: 40**

Answer ALL Questions

[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]

1	An e-commerce company monitors its daily operations. Over one quarter, it records for each day the amount spent on online advertising, the sales generated, and the number of orders received. Across this period, the covariance between advertising spend and sales is $Cov(spend, sales) = 6$, with standard deviations $S_{spend} = 4$ and $S_{sales} = 5$. The covariance between advertising spend and number of orders is $Cov(spend, orders) = 4$, where $S_{orders} = 3$. The covariance between sales and orders is $Cov(sales, orders) = 7$. Separately, the company logs the number of visitors to its website every day for several months. For this daily-visitor series, the lag-1 autocorrelation is $r_1 = 0.8$, and the standard deviation of daily visitors is $S_{orders} = 3$.		**CLO-2 Level-5**
a)	Measure the **correlation coefficient** between **advertising spend** and **sales**, and explain why the analyst should rely on correlation rather than covariance when comparing this company against competitors that trade in different currencies.	[Marks 5]	
b)	Explain what **autocorrelation** measures for the daily-visitor series (including why it carries the prefix "**auto**" and what value ρ_0 must always take), and interpret what $r_1 = 0.8$ reveals about visitor counts on consecutive days.	[Marks 3]	
	The analytics team at a regional energy utility uses short-term weather signals to anticipate electricity demand: on warmer days, air-conditioning load rises, so even a modest ability to guess tomorrow's temperature from today's helps the team schedule generation and avoid buying expensive last-minute power. Before committing to any forecasting model, an analyst runs a quick sanity check on the data; she wants to know whether the temperature series carries useful "memory" (whether one day's reading tells you anything about the next day's), and to summarise the typical level and spread of the readings. She pulls the average daily temperature (°C) recorded by a city weather station over six consecutive days: $y = \{17, 19, 20, 20, 21, 23\}$, $T = 6$		
c)	Determine whether these temperatures show day-to-day persistence: compute the sample mean, the sample variance and standard deviation, form the lag-k	**[Marks 10]**	

		pairs (y_{t-k}, y_t), compute the lag-k sample autocorrelation r_1, and r_2 interpret what its value implies for the analyst's short-term forecasting goal.		
	d)	Draw a simple **neural network** showing the input layer, hidden layer, output layer, weights, biases, and activation functions. Also, **explain why activation functions are necessary** in a neural network.	[Marks 7]	
2	A large healthcare organization operates multiple hospitals and clinics. It collects data from electronic health records, medical imaging systems, laboratory systems, and wearable health devices. Initially, the organization collected data from different sources and sent it to a separate processing system where the data was cleaned, converted into a suitable format, and then stored in a central repository for analysis. This approach worked effectively when the data volume was relatively small. As the organization expanded, the amount and variety of healthcare data increased significantly. Large volumes of patient records, medical images, and monitoring data were generated every day. The separate processing system became slow, causing delays in reporting and decision-making. To improve performance, the organization changed its architecture. Instead of processing data before storage, raw data from different sources was first stored directly in a scalable cloud-based storage platform. Data preparation and analysis tasks were then performed using the computing capabilities of the cloud environment. This allowed the organization to handle larger datasets and support different analytical requirements. Later, the organization introduced a system where data from wearable devices and monitoring equipment was continuously collected, analyzed immediately, and used to generate instant alerts for healthcare providers.			***CLO-3 Level-4***
	a)	Based on the above scenario, **analyse** how the company's data architecture evolved from **ETL to ELT and then to real-time data processing**. Compare the advantages of each stage and discuss the main challenges of maintaining scalability and high performance in real-time data systems.	[Marks 8]	
	A team deploys a machine learning model (say, a fraud detector) through a standard CI/CD pipeline. For several months no one changes the code, and every CI/CD run passes its tests cleanly yet the model's accuracy in production steadily declines and complaints pile up. The engineers are baffled: nothing in the codebase changed, so why is the deployed system getting worse?			
	b)	Explain why a standard CI/CD pipeline, on its own, is not sufficient for a machine learning system, and why MLOps adds **Continuous Training** (**CT**). In your answer, **identify the three things** that can change in an ML system (compared with only one in traditional software) and describe what CT does to address them.	[Marks 7]	