Kasarani Campus
Off Thika Road
P. O. Box 49274, 00101
NAIROBI
Westlands Campus
Pamstech House
Woodvale Grove
Tel. 4442212
Fax: 4444175

KIRIRI WOMENS' UNIVERSITY OF SCIENCE AND TECHNOLOGY
UNIVERSITY EXAMINATIONS, 2024/2025 ACADEMIC YEAR
FIRST YEAR, FIRST SEMESTER EXAMINATION
FOR THE DIPLOMA IN SOFTWARE ENGINEERING

DSE 1009: DATA SCIENCE AND BUSINESS ANALYTICS

DATE: 13TH DECEMBER, 2024
TIME: 8:30AM-10:30 AM

INSTRUCTIONS TO CANDIDATES
ANSWER QUESTION ONE (COMPULSORY) AND ANY OTHER TWO QUESTIONS

QUESTION ONE: COMPULSORY (30 MARKS)

a) State the primary objective of business analytics in a business context **(2 Marks)**
b) Compare data science and business intelligence in terms of purpose, methodologies, and typical tools used. **(3 Marks)**
c) Differentiate between structured and unstructured data, providing an example of each type **(3 Marks)**
d) Consider a company that collects data from internal ERP and CRM systems. Explain how these data sources contribute to understanding business performance and customer relations. **(3 Marks)**
e) Imagine you are tasked with analyzing customer feedback from social media and product usage data from an ERP system. Discuss the challenges you might face due to the difference in data structure and how you would approach these challenges. **(6 Marks)**
f) List two common public datasets that data scientists might use for business insights and briefly describe the type of data each provides. **(2 Marks)**
g) State the benefits of using APIs for data collection in data science. **(3 Marks)**
h) List and describe three key elements of effective data storytelling **(3 Marks)**
i) Describe three key benefits of using cloud storage for businesses **(3 Marks)**
j) State TWO ways in which data accountability contributes to data privacy and compliance in an organization **(2 Marks)**

QUESTION TWO: (20 MARKS)

Tumaini Stores is a mid-sized e-commerce company that collects data from various sources, including customer orders, website analytics, and social media. Recently, they noticed inaccuracies in their data reports, with missing customer demographic information, inconsistencies in product prices, and duplicated records. You were recently hired as an intern data analyst to clean and transform the data to ensure high-quality reports for business decisions. As the analyst, your tasks include understanding ETL tools and techniques, improving data quality, and applying transformations.

a) Tumaini Stores customer demographics dataset contains missing values in the "Age" and "Country" columns. Explain the steps you would take to handle these missing values to improve data quality **(3 Marks)**
b) Tumaini Stores has discovered that some product prices are unusually high or low due to data entry errors, which could impact decision-making. Outline any THREE techniques you would use to identify and address these outliers **(3 Marks)**
c) Explain why addressing outliers is crucial for Tumaini Stores **(2 Marks)**

d) The current customer dataset at Tumaini Stores has inconsistent formats for country codes (e.g., "US" vs. "USA") and customer ID encoding. Critically assess which data transformation techniques would be most suitable for standardizing these formats. Justify your choices **(4 Marks)**

e) Tumaini Stores is considering whether to handle missing values in their data by deletion or imputation. Assess the pros and cons of each method, and recommend the best approach **(5 Marks)**

f) Explain how data cleaning specifically benefits the quality of the insights that Tumaini Stores could gain from their analytics **(3 Marks)**

QUESTION THREE: (20 MARKS)

A healthcare startup uses large datasets collected from hospitals, wearables, and health apps to develop personalized treatment plans for patients. They also leverage predictive analytics to identify potential health risks for patients. As the business grows, they face challenges regarding the ethical use of data and ensuring compliance with data governance standards.

a) Identify **THREE** main ethical concerns related to the use of big data in healthcare **(3 Marks)**

b) State how these concerns identified in (a) above apply to the healthcare startup in the case study **(3 Marks)**

c) The startup uses predictive analytics to identify patients at risk of developing chronic diseases. Discuss the potential ethical dilemmas this might create, particularly in terms of bias and fairness **(2 Marks)**

d) Explain the role of data in creating competitive advantages for data-driven businesses like the healthcare startup in the case study **(3 Marks)**

e) State **THREE** potential risks the startup might face when relying heavily on data for business decisions **(3 Marks)**

f) As an entrepreneur in a data-driven business

i. Identify **ONE** importance of data monetization **(1 Mark)**

ii. State **ONE** ethical challenge that could arise from monetizing patient data in the case study **(2 Marks)**

g) As the healthcare startup expands, they want to develop a new product using data from wearable devices. Describe the steps they should take to identify a viable product-market fit for this new offering **(3 Marks)**

QUESTION FOUR (20 MARKS)

A real estate company collects data on various features of houses, such as the number of bedrooms, square footage, location, and proximity to amenities. They want to predict the price of houses in different neighborhoods based on this data. The company is also interested in identifying patterns in customer preferences for types of homes and using these insights for market segmentation.

a) Using examples from the case study distinguish between

i. Supervised and unsupervised learning **(3 Marks)**

ii. Classification and clustering **(3 Marks)**

b) In the real estate market, customers often prefer houses with certain features (e.g., proximity to schools, large backyard). Explain how unsupervised learning methods could be useful in identifying customer segments based on preferences. **(2 Marks)**

c) Suppose the real estate company has a dataset containing historical house prices and their features (e.g., size, location). Using an appropriate supervised learning algorithm explain how you would approach this problem **(3 Marks)**

d) The real estate company wants to classify houses into "high demand" and "low demand" categories based on historical sales data. Describe how you would use logistic regression for this task **(3 Marks)**

e) The company also wants to group customers into different segments based on their preferences for house features. The manager suggests the adoption of K-Means or hierarchical clustering techniques. Advice him/her on which of the two to select and why **(3 Marks)**

f) After training a linear regression model, you find that the MAE is 10,000 and the RMSE is 15,000.
   i. State what these metrics tell you about the model's prediction errors **(2 Marks)**
   ii. identify which one would you prioritize **(1 Mark)**

QUESTION FIVE (20 MARKS)

Jumia an online retail platform tracks customer behavior and purchase data. They've collected data on the likelihood of customers making a purchase based on visiting the website and the promotional activities they interacted with. Additionally, the company is analyzing the satisfaction scores of customers to determine whether a new marketing strategy impacts overall customer satisfaction. The marketing team wants to understand customer behavior better by using probability, statistical measures, and hypothesis testing.

a) Using examples from the case study above distinguish between
   i. basic probability and conditional probability **(3 Marks)**
   ii. independent and dependent events **(3 Marks)**

b) Suppose the probability of a customer making a purchase after visiting the website is 0.3. The probability of a customer interacting with a promotional activity is 0.6. Given that a customer interacted with a promotional activity, the probability of making a purchase rises to 0.5.
   Using the case study data, calculate:The probability that a randomly selected customer who visited the website and interacted with a promotion made a purchase **(2 Marks)**

c) From the case study, Jumia has satisfaction scores of 100 customers before and after implementing the new marketing strategy. The pre-strategy satisfaction score mean is 70 with a standard deviation of 8. The post-strategy mean is 75 with a standard deviation of 7. Calculate the variance for both datasets **(2 Marks)**

d) State a null hypothesis and an alternative hypothesis in the context of Jumia's goal of testing whether the new marketing strategy improved customer satisfaction **(2 Marks)**

e) Identify ONE scenario where Jumia may incorporate Bayes theorem **(2 Marks)**

f) The marketing team gathers data from two different customer demographics, showing a higher standard deviation in satisfaction scores for one demographic. Discuss what a higher standard deviation means in this context and its implications for future marketing strategies. **(3 Marks)**

g) A competitor uses a different marketing strategy and claims to have a 20% lower variance in customer satisfaction. Based on the statistical measures of variance, discuss whether a lower variance is always beneficial in customer satisfaction analysis. **(3 Marks)**