UGANDA INSTITUTE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY
PROGRESSIVE ASSESSMENT EXAMINATIONS

ACADEMIC YEAR 2025/2026

PROGRAMME(S): DIPLOMA IN DATA SCIENCE MANAGEMENT AND ANALYTICS (DSMA)

YEAR OF STUDY: ONE SEMESTER TWO

COURSE: R PROGRAMMING

COURSE CODE : DSA1206

DATE: FRIDAY 26TH, JUNE 2026

TIME: 9:00 AM – 12:00 PM

DURATION: 3 HOURS

INSTRUCTION:
This paper consists of two sections A and B
Attempt two Questions from section each section.
Create a folder on the desktop in your names and registration number where you will save your work.

SECTION A (40 marks)

Question one

(a) A data frame and a matrix are both two-dimensional structures in R, but they serve fundamentally different purposes.

(i) Explain the primary structural difference between a **matrix** and a **data frame**. *(4 Marks)*
(ii) Provide a practical scenario or dataset example where a data frame must be used instead of a matrix. Explain why. *(4 Marks)*

(b) Exploratory data analysis is one of the crucial steps in any data science project. Briefly its major role in data science. (2 marks)
(c) Differentiate between summary() and str() functions as used in Exploratory Data Analysis (2 marks)
(d) State the use of following functions in Exploratory data Analysis in R.

(i) is.na() (1 marks)
(ii) na.ommit() (1 marks)

(e) state the use of the following functions as used in dataframe plier (dplyr)

(i) select() (2 marks)
(ii) filter() (2 marks)
(iii) arrange() (2 marks)

Question two

(a) Distinguish between categorical data and Outliers as used in data analysis. (2 marks)
(b) The R code below contains syntax errors. Identify the lines where the errors are found and the specify the error on the particular line. Then re-write the correct error-free code (6 marks)

```
1  First_name<-readline("Enter your first name: )
2  Age<-as.numeric(readline(("Enter your age:")))
3  if (Age<18){
4    print("You are a minor")
5  }
6  else if (Age==18){
7    print("You have jumped out of minors")
8  }
9  else{
10   print(paste(First_name, "you are no longer a child. Behave maturely"))
11 }
```

(c) When working with real world data, sometimes data is scattered in different tables and one needs to put them together using a common identifier. This is achieved using R joins. Describe three types of joins used in R. (6 marks)
(d) Distinguish between lists and factors as used in R programming. (2 marks)

(e) Alice is a student aged 24 years and obtained 56, 70, 80 in three different course units. Create a list to store the records about Alice. (4 marks)

Question three

(a) Define the term data visualization (2 marks)
(b) Outline any five importance of visualizing data. (5 marks)
(c) Outline the principles of a good visualization (5 marks)
(d) Ggplot2 is one of the frequently used libraries for data visualization in R. Write the syntax of the ggplot() function (2 marks)
(e) State any three data visualization charts used in R and their respective use cases. (6 marks)

SECTION B (60 marks)

Question four

You have been provided with a large e-commerce dataset named products.csv containing information on different retail products sold online.

(a) .
- (i) Use R code to import the dataset file into a data frame variable named inventory_data. *(2 Marks)*
- (ii) Write the R code to identify the columns present in the dataset and their respective data types. (2 marks)
- (iii) Write the code to check for any rows containing missing values (NA) within inventory_data, remove those incomplete rows, and save the clean dataset into a new data frame called cleaned_inventory. *(4 Marks)*

(b) An analyst wants to deduce consumer demand per product category based on warehouse depletion. Lower stock levels indicate high sales velocity, while high stock levels indicate stagnant inventory.

Write an R script using a for loop combined with an if-else statement to iterate through the sample_stock vector. For each stock level, the script must print a specific message based on the following warehouse thresholds:

- If the stock is **less than 20**, Critical Restock Required!"
- If the stock is **between 20 and 100 (inclusive)**, Low Stock warning."
- If the stock is **greater than 100**, Stock Level Adequate." *(8 Marks)*

(c) Management wants to classify these products into three distinct price tiers to help optimize marketing strategies: **Cheap:** Any product costing strictly less than 50 units, **Moderate:** Any product costing from 50 units up to strictly less than 200 units. **Expensive:** Any product costing 200 units or more.

Management also wants to know the average price per product category and maximum stock per category.

- (i) Create a new column within the data frame named Price_Category. Use nested conditional logic (ifelse) to automatically assign the categories "Cheap", "Moderate", or "Expensive" to each record based on the thresholds specified by management. (4marks)

(ii) Write the command to generate and print a frequency distribution table of the newly created Price_Category column to the R console so management can see the total count for each tier. (2 marks)
(iii) Write the R code using the tapply() function to calculate the average price per product category. (2 marks)
(iv) Write the code to calculate the maximum stock per category.

(d) Plot bar graph showing the distribution of product category. (4 marks)
(e) Write the R command to export this finalized cleaned_inventory data frame to a new spreadsheet file named inventory_promo_report.csv. *(2 Marks)*

Question five

The built-in R dataset msleep contains sleep and feeding data for different species of mammals. The dataset includes missing values due to the difficulty of gathering physiological data for wild animals. Some of the columns include

- name: Common name of the mammal
- vore: Feeding behavior, e.g., "herbi", "carni", "omni"
- sleep_total: Total number of hours slept per day
- brainwt: Brain weight of the mammal in kilograms

(a) Use R code to load the tidyverse/ggplot2 library and assign the msleep dataset to a new data frame variable named mammal_data. *(2 Marks)*
(b) Missing data in biological metrics is common. Write the R code to filter mammal_data so it only keeps rows where the vore column is **not** missing (NA). Save this filtered dataset into a new data frame called cleaned_mammals. *(4 Marks)*
(c) Write a complete R script to accomplish the following data manipulation and plotting tasks:

(i) Load the ggplot2 library. (1 Mark)
(ii) Filter the msleep dataset to remove all rows where the vore column contains missing values (NA). Save this filtered dataset into a new variable called clean_msleep. *(3 Marks)*
(iii) Using the clean_msleep data frame, write the ggplot2 code to construct a Boxplot that maps vore on the x-axis and sleep_total on the y-axis. (3 Marks)
(iv) Add clear, professional titles to your plot: set the main title to "Mammalian Sleep Distribution by Feeding Behavior", label the x-axis as "Feeding Behavior (Vore)", and label the y-axis as "Total Daily Sleep (Hours)". *(3 Marks)*

(d) Neurologists often look at the relationship between body metrics and sleep depth.

(i) Write a user-defined function named calculate_sleep_index that takes two numeric arguments: sleep_hours and brain_weight. The function should calculate and return a "Sleep Index" based on these custom rules:
  - If the brain_weight is missing (is.na), the function should return a default index value of 0.
  - If the brain_weight is greater than 0.1 kg, the function should multiply sleep_hours by 1.5 and return the result.
  - If the brain_weight is 0.1 kg or less, the function should multiply sleep_hours by 2.0 and return the result. (7 Marks)

(ii) Show how you would apply your calculate_sleep_index function to the sleep_total and brainwt columns of the cleaned_mammals data frame to create a brand new column named sleep_index. *(3 Marks)*

(iii) Write the R command to export this completed cleaned_mammals data frame to a CSV file named mammal_sleep_analysis.csv. *(2 Marks)*

Question six

You have been provided with a dataset named `movies.csv` containing historical metadata about films (including titles, genres, budgets, gross earnings, release years, and audience scores). As a Data Scientist, your task is to clean this data, extract meaningful insights, and visualize trends to help a production company understand what makes a movie financially successful.

(a) Load the dataset into your environment and perform the following structural clean-up operations

(i) Identify columns with missing (`NaN` or `null`) values. Drop rows where critical financial information (`budget` or `gross`) is missing. For missing `rating` or `country` values, fill them with the string `"Unknown"`. *(4 Marks)*

(ii) Check the data types of your columns. Convert the `budget` and `gross` columns from float/object to integers, and convert the `released` column to a standard datetime format. *(4 Marks)*

(b) .

(i) Create a new column named `profit` by calculating the difference between `gross` and `budget` (`gross - budget`). *(3 Marks)*

(ii) Extract the **year** of release as a new standalone column named `release_year` directly from your cleaned datetime column. *(3 Marks)*

(c) .

(i) Calculate the average budget and average profit for each movie `genre`. Sort the results in descending order of profitability. *(4 Marks)*

(ii) Identify the top 5 highest-grossing movies released between the years 2010 and 2020 (inclusive). Display only their `title`, `genre`, and `gross` earnings. *(4 Marks)*

(d) .

(i) Create a line plot showing the trend of average movie budgets changing over the years (from the oldest year to the newest year in the dataset). *(4 Marks)*

(ii) Create a bar chart comparing the total or average profit across the top 5 most frequent genres in the dataset. *(4 Marks)*

(iii) Export the images of the visualizations and save them in the folder containing your work. (2 marks)

END