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, SECOND SEMESTER EXAMINATION
FOR MASTER OF SCIENCE IN APPLIED STATISTICS AND DATA ANALYTICS
KMA 5114: PRINCIPLES OF DATA SCIENCE

DATE: 29TH JANUARY, 2025
TIME: 9:00 AM – 12:00 PM

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

QUESTION ONE: COMPULSORY (40 MARKS)

a) Briefly explain any four data objects used in R Statistical Software. **(4 Marks)**
b) Brief, explain each of the following Machine Learning Methods:
   i) Decision Trees. **(2 Marks)**
   ii) Random forest. **(2 Marks)**
   iii) Principal Component Analysis (PCA). **(2 Marks)**
   iv) Logistic Regression **(2 Marks)**
c) What do you understand by the term **Big Data** giving its characteristics? **(5 Marks)**
d) Suppose 5 students in second year class got the following scores in their final exam 85, 75, 58, 65, 78. Write an R program to:
   i) Calculate the average scores. **(2 Marks)**
   ii) Suppose 3 more students have their scores as 72, 65 and 59 combine them with the previous ones and determine their mean squared deviations. **(2 Marks)**
   iii) Suppose a big data of 100,000 values is give would your codes change and how? **(2 Marks)**
e) State the two common data formats and explain them briefly. **(4 Marks)**
f) What is Data visualization? Give examples of how data is visualized. **(4 Marks)**
g) Wafula wanted to buy a mobile phone, and he found that the prices ranges from different shops 15000, 19400, 10200, 29000, 30500, 20000, 49500 (in Ksh)
   Write an R program that
   i) Input the data into R with a vector name PRICE and checks the number of entries. **(3 Marks)**
   ii) Suppose 10,200 was a mistake, it should have been 20,800. How can you fix this? **(3 Marks)**
   iii) State the output of the following R codes: PRICE[5], PRICE[2:6], PRICE[-4]. **(3 Marks)**

QUESTIONS TWO: (20 MARKS)

a) Artificial neural networks (ANNs) are model from biological neurons. Outline the biological properties that are borrowed into ANNs. **(4 Marks)**

b) Define what is Perceptron and state its two types. **(3 Marks)**

c) Explain what is forward and backpropagation in Artificial neural networks (ANNs). **(4 Marks)**

d) Distinguish between a simple ANN and the deep learning approach. **(4 Marks)**

e) You have been hired as the data scientist in the meteorological department. Your first assignment involves the use of neural networks for time series prediction specifically weather forecasting. Explain the techniques that you would apply for this task and justify why. **(5 Marks)**

QUESTIONS THREE: (20 MARKS)

a) What do you understand by Ethics in Data Science? also explain what is meant by data snooping **(4 Marks)**

b) Given the following data set

Set 1: 1, 2, 3, 2, 1, 1, 6, 4, 4, 7, 2, 5, 10, 2, 0

Set2: 1, 3, 5, 2, 0, 1, 3, 4, 2, 4, 7, 3, 1, 5, 1, 2

Write an R program that Creates the following two matrices A and B each having four rows, arranging the elements row wise. **(6 Marks)**

c) Interpret the following R Code. **(10 Marks)**

```
reg3<-lm(prestige~education+log(income)+type,data=Prestige)
summary(reg3)
Call:
lm(formula = prestige ~ education + log(income) + type, data = Prestige)
Residuals:
    Min      1Q  Median      3Q     Max
-13.511  -3.746   1.011   4.356  18.438
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -82.6413    13.7875  -5.994 3.86e-08 ***
education     3.2845     0.6081   5.401 5.06e-07 ***
log(income)  10.4875     1.7167   6.109 2.31e-08 ***
typebc        1.4394     2.3780   0.605   0.5465
typeprof      8.1903     2.5882   3.165   0.0021 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.637 on 93 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.8555,    Adjusted R-squared:  0.8493
F-statistic: 137.6 on 4 and 93 DF,  p-value: < 2.2e-16
```

QUESTIONS FOUR: (20 MARKS)

a) What is Data visualization? Give examples of how data is visualized. **(4 Marks)**

b) Write a R code on how to create a scatter plot and Boxplot. **(6 Marks)**

c) This is an open data set by Uniform Crime Reporting Statistics consisting of 50 measurements of 7 variables. It states for 1 year (2021) the reported number of crimes in the 50 states of the U.S. classified according to 7 categories (X3–X9) as stated below:

X1: land area (land),
X2: population 1985 (popu 1985),
X3: murder (murd),
X4: rape,
X5: robbery (robb),
X6: assault (assa),
X7: burglary (burg),
X8: larcery (larc),
X9: autothieft (auto),
X10: U.S. states region number (reg), and
X11: U.S. states division number (div).

Write an R code that would be used to:

i) Help to visualize the data **(3 Marks)**

ii) Classify the data. **(2 Marks)**

iii) Give the most committed crime, in which state by which age and gender. **(5 Marks)**

QUESTION FIVE (20 MARKS)

a) Explain briefly what each of the following loop in R programming does as you write their syntax.
(i) for ( ) (ii) if ( ) (iii) while ( ) **(6 Marks)**

b) Write an R-program using loop that would solve the following equation. **(7 Marks)**

$$\sum_{x=-1}^{3}\sum_{y=1}^{5}\sum_{z=0}^{6}\left(\frac{xyz^2}{4+5x-3z}\right)$$

c) The weight of 50 students from Kiriri university was recorded as follows

68 73 80 56 38 81 36 64 36 82 38 58 78 80 73 57 68 77 86 81 63 39 51 87 37
82 68 72 85 88 47 66 88 49 37 47 79 39 68 60 36 36 64 44 47 38 56 80 44 79

Jane wanted to classified the weights as follows

Class	Status
Below 40 KG	Under
$40 \leq \times < 55KG$	Feather
$55 \leq \times < 70KG$	Middle
70KG and above	Heavy

Write an R-code using a nested loop that tests the weights of each student and give the correct status. **(7 Marks)**