MERU UNIVERSITY OF SCIENCE AND TECHNOLOGY
P.O. Box 972-60200 – Meru-Kenya
Tel: +254(0) 799 529 958, +254(0) 799 529 959, + 254 (0) 712 524 293,
Website: info@must.ac.ke Email: info@must.ac.ke

University Examinations 2023/2024

FOURTH YEAR SECOND SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF DATA SCIENCE

CDS 3450: EXPLANATORY DATA ANALYSIS

DATE: APRIL 2024 **TIME: 2 HOURS**

INSTRUCTIONS: *Answer question **one** and any other **two** questions*

QUESTION ONE (30 MARKS)

a) Using a suitable example; differentiate between the following terminologies as used in exploratory data analysis

i. Univariate Analysis (2marks)

ii. Multivariate Analysis (2marks)

iii. Time Series Analysis (2marks)

b) Unilіver company want to enhance its marketing campaigns within Meru region to increase customer engagement and drive sales for their new soap product. Briefly explain how the company can optimize their marketing strategies using EDA. (6 marks)

c) Consider the dataset given as follows.

Class interval	10-20	20-30	30-40	40-50	50-60
Frequency	5	8	12	16	10

i. Find the mode (4 marks)

ii. When is the Mode a good measure of Central Tendency? (2 marks)

iii. Briefly explain the purpose of central tendency? (2 marks)

d) Highlight key steps involved in the data collection process in a data science project at Naivas Supermarket (6 marks)

e) Discuss any two common challenges encountered during the data cleaning process, and how can they be addressed (4 marks)

QUESTION TWO (20 MARKS)

a) Briefly discuss how version control using Git and GitHub benefit collaborative data science projects (6 marks)

b) Data Visualization is the process of analyzing data in the form of graphs or maps. What is the role of data visualization in data analysis within Meru University? (6 marks)

c) Matplotlib and Seaborn are python libraries that are used for data visualization. Briefly distinguish between the two libraries (4 marks)

d) Highlight two advantages of using interactive data visualization tools over static ones? (4 marks)

QUESTION THREE (20 MARKS)

a) Discuss how dependency relationships between variables influence data analysis and modelling in exploratory data analysis? (6 marks)

b) The following code illustrate relationship with categorical features. Translate the code giving the expected output. (6 marks)

```
#box plot overallqual/saleprice
var 'OverallQual'
data =   pd.concat([df_train[‘SalePrice’],df__train[var]], axis=l)
f, ax =plt.subplots(figsize=(8, 6))
fig sns.boxplot(x=var, "SalePrice", data=data)
fig.axis(ymin=0, ymax=800000);
```

c) Discuss two techniques for handling multivariate categorical variables in data analysis. (4 marks)

d) Highlight two advantages of using box plots over histogram (4 marks)

QUESTION FOUR (20 MARKS)

a) Consider a unvaried data set $x = x_1 .....x_n$. Write down commands in python to compute statistical moments. (5 marks)

Further write commands to plot histogram, bar plots and box plots in python (6 marks)

b) How does temporal data differ from other types of data, and what considerations are important when analyzing it? (5 marks)

c) Describe the characteristics of spatial data and explain why it requires specialized analysis techniques. (4 marks)

QUESTION FIVE (20 MARKS)

a) Consider the data below

6.4	6.6	6.2	7.3	6.2	8.1	7.0
7.0	5.9	5.7	7.0	7.4	6.5	6.8
7.0	7.0	6.0	6.3	5.6	6.3	5.8
5.9	7.2	7.3	7.7	6.8	5.2	5.2
6.4	6.7	6.2	7.5	6.8	6.4	7.8

i. Compute Arithmetic mean, Geometric mean and harmonic mean to verify the relationship (4 marks)

ii. Write a python code for implementing the above measures of central tendency in 5(a)i

(7 marks)

b) Briefly explain the role of preprocessing in preparing data for analysis (3 marks)

c) Write a simple code to illustrate how transformation of data can be done in python (6 marks)