Programme Code: TU856
Shared with: TU857
Module Code: CMPU1039
CRN: 31597/31481

TECHNOLOGICAL UNIVERSITY DUBLIN

CITY CAMPUS - GRANGEGORMAN

TU856 BSc. (Honours) Degree in Computer Science

TU857 BSc. (Honours) Degree in Computer Science (Infrastructure)

Year 1

SEMESTER 2 EXAMINATIONS 2023/24

CMPU 1039 Data Exploration

Internal Examiner:
Jane Ferris
Dr. Paul Doyle

Exam Duration: Two Hours

Instructions To Candidates:

Answer Question One AND **Two** Other Questions.

Question One Is **Compulsory** & Carries 50 Marks.

All Other Questions Carry 25 Marks

Special Instructions /Handouts/ Materials Required: None

Question One. This Question is Compulsory.
All subsections have equal marks.

a) Identify the scale of measurement (not the data type) for the following data variables:
   i. 'age'
   ii. 'hair_colour'
   iii. 'qualification'

(5 marks)

b) Rank the scales of measurement data categories according to their level of significance in data analytics.

(5 marks)

c) Identify the best measure to describe the centrality of the data in relation to the data variable ‘hair_colour’.

(5 marks)

d) Is the SE of a data calculation a robust measure?

(5 marks)

e) In a study of ‘strength’ (kgs) and ‘hair_colour’ (blond, brunette, black and other) Identify which statistical test must be performed to assess the relationship between the value of an individual’s strength and their hair colour.

(5 marks)

f) An experiment to test the tensile strength of paper from two different suppliers was carried out. The results are displayed below.

Sample	N	Mean	Median	Standard Deviation
X	15	87.32	86.45	18.32
Y	96	83.45	84.78	1.5

Describe the significance of the standard deviation reported above and provide one *(1)* recommendation for future analysis.

(5 marks)

g) State the Null and Alternate Hypothesis for the claim that there's a significant difference in the tensile strengths of the paper of brand X and brand Y.

(5 marks)

h) If the significance threshold is 0.05 what is the outcome of an inferential test comparing brand X and brand Y that calculates a probability of .23?

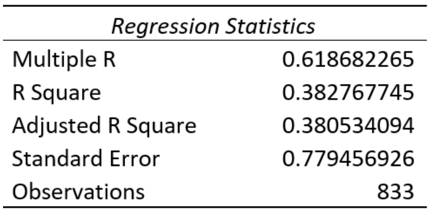
(5 marks)

i) What is the significance of Anscombe quartet in the field of data visualisation?

(5 marks)

Question One continued.
This Question is Compulsory

j) The regression statistics for a data model is reported below, describe the value and significance of the regression determinate coefficient.



Regression Statistics	
Multiple R	0.618682265
R Square	0.382767745
Adjusted R Square	0.380534094
Standard Error	0.779456926
Observations	833

(5 marks)
(Total 50 marks)

Question Two.

a) Identify the purpose of meta data, giving an example of meta data in your answer.

(5 marks)

b) Identify three *(3)* important reasons why Excel is used in preference to Tableau in data analytics.

(5 marks)

c) Identify three *(3)* important reasons why a DB is used in preference to Excel in data analytics.

(5 marks)

d) Identify three *(3)* important advantages of using a DBMS over an equivalent paper based system?

(5 marks)

e) Identify how an individual’s ‘Age’ variable is stored in Access, note any actions required to store the variable correctly.

(5 marks)
(Total 25 marks)

Question Three.

The daily temperatures (degrees Celsius) in Dublin in early February are provided below:

February Daily Temperatures										
Mon. 1st	Tue. 2nd	Wed. 3rd	Thurs. 4th	Fri. 5th	Sat. 6th	Sun. 7th	Mon. 8th	Tue. 9th	Wed. 10th	Thurs. 11th
2	3	6	7	16	5	3	2	3	6	7

a) Which visualisation is the most appropriate for representing the February average daily temperatures and the measured Inter Quartile Range (IQR)?

(5 marks)

b) Using the data provided above, calculate the range and the IQR for the February average daily temperatures. Show all calculation steps.

(5 marks)

c) Explain the difference between the range and IQR measured for the February average daily temperatures and recommend which measure should be used to describe the data.

(5 marks)

d) Match the most appropriate term to the analytic study described below.
"A study of the underlying conditions for the rise in early February temperatures in Ireland."

a. Prescriptive analytics
b. Descriptive analytics
c. Diagnostic analytics

(5 marks)

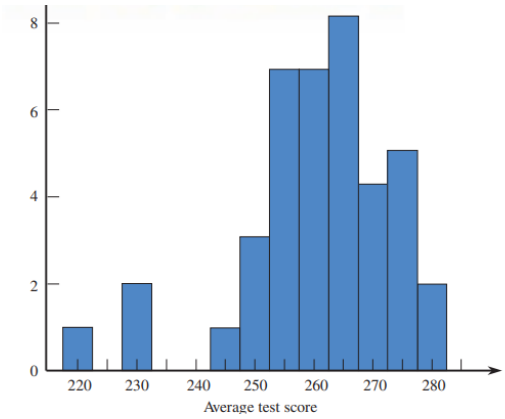
e) A Linear Regression model is applied to forecast the average daily temperatures for late February based on the early February data.
Please identify the bias associated with using this specific machine learning algorithm for this dataset.

(5 marks)
(25 marks)

Question 4 is presented on page 5

Question Four.

You are part of a data exploration team investigating the Test Scores of TUDublin, School of Computer year 1 students.
There are 350 students in year 1.
The mean test score is 260 and the standard deviation is 15 for the year.
Below is a visualisation of the results.



a) With reference to the histogram presented above, identify if this is a sample or a population study.

(3 marks)

b) With reference to the histogram presented above, identify if the distribution above is leptokurtic or platykurtic.

(5 marks)

c) If the team investigates the relationship between the test scores and the amount of time spent studying by students at TUDublin which statistical test should they use?

(3 marks)

d) If the team investigates the students' postal address townland and their test score which statistical test should they use?

(5 marks)

e) A new subject's data is recorded with a value of 220.
How does the new student compare to the Test Scores of the group?
Provide all calculations with your answer in relation to the classification of the subject in relation to the group statistics.

(9 marks)
(Total 25 marks)

End of Paper