Programme Code: TU856/TU857
Module Code: CMPU1039

TECHNOLOGICAL UNIVERSITY DUBLIN

Grangegorman

TU856 BSc. (Honours) Degree in Computer Science

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

Year 1

SEMESTER 2022/23

CMPU 1039 Data Exploration

Internal Examiner(s):
Dr. Paul Doyle
Jane Ferris

Instructions To Candidates:

Answer Question One & Two Other Questions.

Question One Is Compulsory & Carries 50 Marks.

All Other Questions Carry 25 Marks

Exam Duration:

Two Hours

Special Instructions /Handouts/ Materials Required:

None

Question 1. This Question is compulsory.

a) Identify the software associated with the following file formats:
   i) .ACCDB
   ii) .TWB
   iii) .CSV

(5 marks)

b) What is the primary difference between ratio and interval data types? Give an example of both data types.

(5 marks)

c) Big Data is described by the three (3) Vs. Identify what the 3 Vs are.

(5 marks)

d) Describe the *mean*, the *mode* and the *median* give examples of each measure.

(5 marks)

e) Identify the bias that may occur in an online survey in relation to a study to evaluate the financial needs of the parents of TUDublin students.

(5 marks)

f) In an analysis of popcorn salt levels, the following sample statistics are found:

Mean (g/Kg)	Standard Deviation	Standard Error
.13	.25	.09

Describe the value and significance of the standard error reported.

(5 marks)

g) Identify the three most important data cleaning steps in the processing of a MS Excel dataset.

(5 marks)

h) A dataset has been provided to you for analysis in MS Excel with a '?' placeholder for not known values. Identify the method you will use to clean or correct the data and note any issues that you may experience.

(5 marks)

i) Identify three pre-analysis data and data use requirements that relate to the General Data Protection Regulations introduced to Europe in 2016.

(5 marks)

j) Identify three advantages of using a Data Base Management System over the use of MS Excel to electronically store data files.

(5 marks)

(Total 50 marks)

Question Two

The following human resources (personnel) data has been provided to you to analyse and store by the Ferris Company.

Martin P Adams, DoB; 4/5/1976, Personal Public Service Number (PPSN): 876532A, started 5/6/2001, pay 7 euros/hour
Dorothy Burke, DoB; 12/8/1979, Personal Public Service Number: 235149C, started 7/9/2001, pay 6.50 euros/hour
Noel Bourke DoB; 5/8/1969, Personal Public Service Number: 563420Q, started 19/8/1998, pay 8.50 euro/hour
John Clark, DoB; 6/7/1957, Personal Public Service Number: 673401A, started 2/12/1986, pay 15 euros/hour
Noleen Conners, DoB; 12/5/1947, Personal Public Service Number: 543876Z, started 3/2/1985, pay 14.50 euros/hour
Mary Doyle, DoB; 27/3/1981, Personal Public Service Number: 543729S, started 12/12/2001, pay 5 euros/hour

---

a) Draw the 'Personnel' Database Table Entity Relationship Diagram for the data provided. Ensure that you identify the required data formats for storage.

(8 marks)

b Give two (2) constraints on variables that are identified as the Primary Key in a DataBase.

(4 marks)

c What is the Primary Key for the 'Personnel' DataBase Table?

(2 marks)

d Explain the term derived data in relation to the Personnel table.

(4 marks)

e) There is a claim that there's a Gender Pay Gap within the Ferris Company. Answer the three following question in relation to the data claim.

a) Identify which data from the 'Personnel' Table would be used to investigate the claim.

(1 marks)

b) Identify which specific analysis would be required to test this claim.

(3 marks)

c) Identify an error that is inherent in the analysis.

(3 marks)

(Total 25 marks)

Question Three.

a) Summary statistics for the two variables of a breast cancer study are provided below.

UniformCellShape			
Class	Mean	Standard Deviation	Count
2	1.42	.04	450
4	6.51	.16	238
MargAdhesion			
Class	Mean	Standard Deviation	Count
2	1.37	.05	450
4	5.51	.21	238

Write the hypothesis tests that relate to a comparative analysis of the 'UniformCellShape' variable in relation to class '2' and class'4' of the data.

(4 marks)

b) A new subject's data is recorded with a value of 3 in 'UniformCellShape'. In relation to the summary statistics reported in part a), how does the subject relate to the sample class '2'?
Please provide all calculations with your answer in relation to the classification of the subject in relation to the sample.

(6 marks)

c) A cancer diagnostic classifier is built using a statistical analysis technique. The output of the analysis is provided below.

Regression Statistics	
Multiple R	0.842094
R Square	0.709122
Adjusted R Square	0.708274
Standard Error	0.51352
Observations	689

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	441.01	220.505	836.1881	1.1E-184
Residual	686	180.9	0.263703		
Total	688	621.91			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	1.786647	0.029529	60.5058	3E-277	1.72867
1	0.209564	0.009054	23.14542	4.91E-88	0.191786
1	0.083684	0.009376	8.925425	4.02E-18	0.065275

Answer the following four questions in relation to the output provided.

i) What is the Correlation Coefficient?

(2 marks)

ii) What is the Coefficient of the Determinant of the Regression?

(3 marks)

iii) What is the Linear Regression Model that will be used to classify the data?

(5 marks)

iv) What value or values in the output will indicate the accuracy of the performance of the model to correctly classify the cancers?

(5 marks)

(Total 25 marks)

Question Four

You are part of a data exploration team investigating the Titanic dataset in MS Excel.
The following variables have been provided:

1. Passenger Age (years)
2. Passenger Fare (Pounds Sterling 1912 value)
3. The Passenger Sex (Male or Female)
4. The Passenger Survival Status (1 or 0)

a) Identify the three most important data visualisations that are used to explore datasets.

(6 marks)

b) Identify the most appropriate chart for each of the three specific required representations:

i) To summarise the Passenger Sex and the Survival status variables.

(2 marks)

ii) The mean value of the Passenger Age variable.

(2 marks)

iii) The occurrence of outliers in the Passenger Fare variable.

(2 marks)

c) If the team investigate the occurrence of the Passenger Sex variable on board the Titanic with other similar transatlantic passenger ships of the time, which test should they use?

(4 marks)

d) If the team investigate the association between the Passenger Fare and the Passenger Age which test should be used?

(4 marks)

e) A test returns a ‘p’ value of .05.
Identify what the ‘p’ value is and what the value indicates for the test with reference to the standard significance value.

(5 marks)

(Total 25 marks)