Programme Code(s): TU856, TU857
Module Code: CMPU1039
CRN(s): 31481, 31597

TECHNOLOGICAL UNIVERSITY DUBLIN

CITY CAMPUS

TU856 – BSc. (Honours) in Computer Science
TU857 – BSc. (Honours) in Computer Science (Infrastructure)

Year 1

SEMESTER 2 EXAMINATIONS 2021/22

Data Exploration

Ms. Jane Ferris
Dr. Paul Doyle

Instructions

ANSWER *QUESTION ONE* & TWO OTHER QUESTIONS.

QUESTION ONE IS *COMPULSORY* & CARRIES 50 MARKS.

ALL OTHER QUESTIONS CARRY 25 MARKS.

Question 1. This Question is compulsory.

a) Expand the acronyms: DBMS; CSV and ERD

(5 marks)

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

(5 marks)

c) What is the primary difference between a discrete qualitative and a discrete quantitative data type? Give an example of both data types.

(5 marks)

d) What is the difference between .CSV and a .TXT file and give one reason why they are used in preference to a XLSX file for providing your data to other data explorers.

(5 marks)

e) Is linear regression a typical application in data exploration? Please give reasons why.

(5 marks)

f) Describe the *mean* and *median* and identify which are resilient to *outliers*.

(5 marks)

g) What is the difference between data cleaning and data transformation?

(5 marks)

h) What is the correct image to represent the distribution, mean and inter quartile range of data?

(5 marks)

i) Why in the age of big data is inferential statistics still significant in modern data exploration and data analytics?

(5 marks)

j) Identify three advantages of a DataBase over the use of other paper or electronic file stores.

(5 marks)

(Total 50 marks)

Question Two

a) A dataset is received from a study in Tallaght Hospital that details every fracture and strain recorded in 1 month in 2022 in the Emergency Department. Is this a sample or a population dataset? Please give reasons why with your answer.

(5 marks)

b) Give three (3) advantages of using a population in preference to using a sample.

(5 marks)

c) What is Standard Error (SE) and how does it relate the mean of a sample measured to the estimated mean of the population?

(7 marks)

d) Using the summary statistics provided in Appendix A (on the last page of this exam), calculate the SE for the 'Brand A' product and report the estimated population mean.

(8 marks)

(Total 25 marks)

Question Three.

a) Define an outlier in relation to *either* the Inter Quartile Range *or* the Standard Deviation measures of a sample.

(5 marks)

b) Describe the distribution of a normally distributed variable.

(5 marks)

c) In normally distributed data such as 'Brand A' the z score is used to express the relatedness of a new sample to the same data. What statistical measures are used in the calculation of the z score?

(7 marks)

d) A new sample of 'Brand A' has been taken and has a measured value of 11.2. In relation to the normally distributed Brand A's summary statistics reported in appendix A (on the last page of this exam), is the new sample related to the original sample studied? Please provide calculations with your answer in relation to the identification of outliers with or without the use of the z score calculation.

(8 marks)

(Total 25 marks)

Question Four

a) A study relating to the sale of popcorn at the local Cinema is about to take place. The genre of the films, the day and time and the sale of popcorn is to be recorded. Which visualisation or visualisations will the manager provide when presenting his analysis? Please give reasons for each of your answers.

(5 marks)

b) Identify the type of *bias* described by the L'Oreal advertisement for moisturiser: "*71% of Women agree that this product reduces visible lines and wrinkles within an hour (based on a survey of 48 women run by Heat Magazine in June, 2020)*".

(5 marks)

c) If a study with a Null Hypothesis states that there is no association between the Zodiac Star Sign of an individual and their Criminality, returns a probability value of .9123. Will this statistically prove that they are independent or dependent variables? Please give reasons why with your answer.

(7 marks)

d) What is 'p hacking': describe the context of use and purpose in inferential statistics.

(8 marks)

(Total 25 marks)

Appendix A on final page.

Appendix A

Brand A	
Mean	9.6
Median	9.5
Mode	10
Standard Deviation	0.6
Sample Variance	0.4
Kurtosis	0.6
Skewness	0.9
Range	2
Minimum	8.9
Maximum	10.9
Sum	77
Count	9