FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

MID-SEMESTER ASSESSMENT

MODULE CODE: MA4128

SEMESTER: Spring

MODULE TITLE: Advanced Data Modeling

DURATION OF EXAM: 1 hour

LECTURER: Mr. Kevin O'Brien

GRADING SCHEME: 100 marks
20% of module grade

INSTRUCTIONS TO CANDIDATES

Scientific calculators approved by the University of Limerick can be used.
Formula sheet and statistical tables provided at the end of the exam paper.
Students must attempt any 4 questions from 5.

Question 1. (10 marks) Distriutional Assumptions

(a) (10 Marks)

The data set `X` and `Y` are both assumed to be normally distributed. The Shapiro-Wilk test was carried out to assess whether or not this assumption is valid for data set `X`.

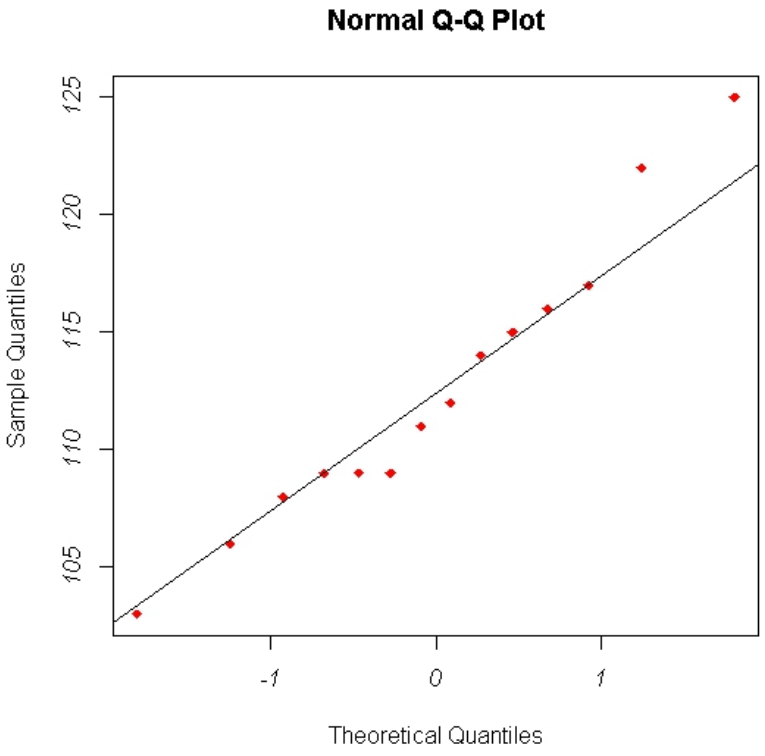
i. (1 marks) Formally state the null and alternative hypothesis.

ii. (2 marks) What is your conclusion for this procedure? Justify your answer.

```
> shapiro.test(X)

Shapiro-Wilk normality test
data:  X
W = 0.9292, p-value = 0.372
```

(c) The data set `X` and `Y` are both assumed to be normally distributed. A graphical procedure was carried out to assess whether or not this assumption of normality is valid for data set `Y`. Consider the Q-Q plot in the figure below.



i. (2 marks) Provide a brief description on how to interpret this plot.

ii. (1 marks) What is your conclusion for this procedure? Justify your answer.

(b) The typing speeds for one group of 12 Engineering students were recorded both at the beginning of year 1 of their studies. The results (in words per minute) are given below:

149	146	122	142	153
137	161	156	170	159

Use the Dixon Q-test to determine if there is an outlier present in this data. You may assume a significance level of 5%.

(i.) (1 Mark) State the null and alternative Hypothesis for this test.

(ii.) (2 Marks) Compute the test statistic

(iii.) (1 Mark) State the appropriate critical value.

(iv.) (1 Mark) What is your conclusion to this procedure.

Question 2. (10 marks) Binary Classification

(a) ***Binary Classification (6 Marks)***
For following binary classification outcome table, calculate the following appraisal metrics.

(i) (1 Mark) accuracy;

(ii) (1 Mark) recall;

(iii) (1 Mark) precision;

(iv) (1 Mark) F-measure.

	Predict Negative	Predict Positive
Observed Negative	9530	10
Observed Positive	300	160

(v) (2 Marks) Explain why the F-measure is considered a more informative measure of performance than the Accuracy score.

(iii.) (2 Marks) Define Specificity and Sensitivity. You make reference to previous answers.

(iv.) (3 Marks) What is a ROC curve? Explain its function, how it is determined, and the means of interpreting the curve. Support your answer with a sketch.

Question 3. (10 marks) Hierarchical Clustering

i. (2 Marks) What is the purpose of a cluster analysis?

ii. (2 Marks) A discriminant analysis is similar to a cluster analysis; however, there is one fundamental difference. Explain this difference.

iii. (2 Marks) What is the difference between a linkage method and a distance measure?

iv. (2 Marks) Compare and contrast any two linkage methods.

- Why do you standardize variables before carrying out a cluster analysis.

  Explain why using the standardized value may not be suitable in some cases?

v. (2 Marks) How do we determine the appropriate number of clusters? Give two different visualization methods that are used to display the outcome of a cluster analysis.

vi. (2 Marks) Standardization

vii. (2 Marks) Explain the difference between Ward's method and k-means clustering.

viii. (2 Marks) In the context of hierarchical cluster analysis, distinguish between agglomerative clustering and divisive clustering.

ix. (2 Marks) What is a vertical icicle plot used for? Give a brief description, supporting your answer with sketches.

x. (2 Marks) Compute the Euclidean distance between the following points.

$$A = (4, 6, 8, 2)$$

$$B = (3, 6, 1, 6)$$

Question 4. (10 marks) K-Means Clustering

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Net profit	495,145	3	1419,744	3	,349	,795
Own funds	2878,202	3	2537,200	3	1,134	,460
Assets	842788,443	3	9987,138	3	84,387	,002
Client deposits	634017,636	3	35643,498	3	17,788	,021
Loans	957411,333	3	37401,709	3	25,598	,012

Figure 1:

Question 5. (10 marks) Modelling Count Variables

(i)

(ii)

(iii)

(iv)

(v)

Formulas and Tables

Critical Values for Dixon Q Test

N	$\alpha = 0.10$ *Confidence*= 0.90	$\alpha = 0.05$ *Confidence*= 0.95	$\alpha = 0.01$ *Confidence*= 0.99
3	0.941	0.97	0.994
4	0.765	0.829	0.926
5	0.642	0.71	0.821
6	0.56	0.625	0.74
7	0.507	0.568	0.68
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
11	0.392	0.444	0.542
12	0.376	0.426	0.522
13	0.361	0.41	0.503
14	0.349	0.396	0.488
15	0.338	0.384	0.475
16	0.329	0.374	0.463