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: 50 marks
20% of module grade

INSTRUCTIONS TO CANDIDATES

Scientific calculators approved by the University of Limerick can be used.
Statistical tables provided at the end of the exam paper.
Students must attempt ALL questions

Question 1. (10 marks) Distributional Assumptions

(a) (5 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`.

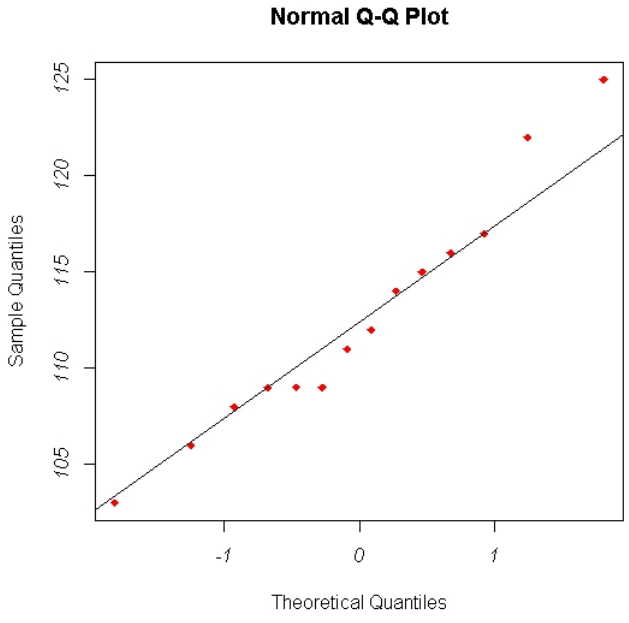
(i.) (1 Mark) Formally state the null and alternative hypothesis.

(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
```

Continuing with the data sets `X` and `Y`, 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.



(iii.) (1 Mark) Provide a brief description on how to interpret this plot.

(iv.) (1 Mark) What is your conclusion for this procedure? Justify your answer.

(b) (5 Marks)

The typing speeds for one group of 10 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	151	156	170	149

Use the Dixon Q-test to determine if there is an outlier present in this data. You may assume a significance level of 5%. Critical values are tabulated at the back of this exam paper.

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

(ii.) (1 Mark) Compute the test statistic

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

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

(v.) (1 Mark) Briefly discuss any limitations upon using this test.

Question 2. (10 marks) Binary Classification

(a) ***Performance Metrics (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.

(b) ***ROC Curves (4 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 sketchs.

Question 3. (10 marks) Hierarchical Clustering

(i.) (1 Mark) Compute the Euclidean distance between the following points.

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

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

(ii.) (2 Marks) Why do you standardize variables before carrying out a cluster analysis. Explain why using the standardized value may not be suitable in some cases? Give another example of numeric transformation.

(iii.) (4 Marks) Compare and contrast any three linkage methods. Support your answer with sketches

(iv.) (1 Mark) In the context of hierarchical cluster analysis, distinguish between agglomerative clustering and divisive clustering.

(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.

Question 4. (10 marks) K-Means Clustering

(i.) (5 Marks) Explain the process of k-means clustering, starting with initial cluster allocation. You may work on the basis of a two-cluster solution. Support your answer with several sketches.

(ii.) (2 Marks) Compare and contrast k-means clustering and hierarchical clustering in terms of the number of cluster determined.

(iii.) (3 Marks) For a 4 cluster solution, Interpret the ANOVA table below.

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

Question 5. (10 marks) Modelling Count Variables

(i.) (2 Marks) What is Poisson regression used to model. State any assumptions that must be checked before it can be used as an analysis.

(ii.) (1 Mark) The `R` Code output given below is used to predict the number of awards won by students.

- Information is provided on which of the three school programs the student takes part in (*General*, *Vocational* or *Academic*).
- Also we are given the mathematics test score.

State the mathematical formula used to predict the numer of awards won. *You can denote* ***progAcademic***, ***progVocational*** *and* ***math*** *as* x_1,x_2 *and* x_3 *respectively.*

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.2471     0.6585   -7.97  1.6e-15 ***
progAcademic     1.0839     0.3583    3.03   0.0025 **
progVocational   0.3698     0.4411    0.84   0.4018
math             0.0702     0.0106    6.62  3.6e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(iii.) (2 Marks) Use the model in Part (ii) to predict the number ofawards won by a general program student, witha maths score of 60.

(iv.) (3 Marks) What is Zero Inflation? Explain the Modelling Process for a Zero Inflated Model. Give an Example of Zero-Inflated Count Process. *Support your answer with a sketch, if necessary.*

(v.) (1 Mark) Describe a situation whereby Negative Binomial Regression Models would be used instead of Poisson Models.

(vi.) (1 Mark) What is Zero Truncation? Give an example of a Zero Truncated Count Process

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