Dublin Business School
excellence through learning

QQI

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

REPEAT EXAMINATIONS

Module Code:	**B8IT109**
Module Description:	**Advanced Data Analytics**
Examiner:	**Dr Shahram Azizi Sazi**
Internal Moderator:	**Mr Darren Redmond**
External Examiner:	**Brett Becker**

Date: 06 November 2018
Time: 18:30 – 20:30

INSTRUCTIONS TO CANDIDATES

1. *It is required to solve all questions with R. R code and necessary outputs need to be saved in the .R format. Please submit the R code and the required graphs/plots/curves in the zipped folder.*
2. *Please select only one question between questions **3 and 4**. Explicitly mention your optional question on the top of exam paper.*
3. *In the case, where the explanation of your output is in R, it is required to mention explicitly on the paper exam. Notice that your explanation should be consistent with your output. This*

exam assesses you on the development of the advanced techniques for data analytics mainly covering:

- ***Descriptive analysis***
- ***probability models***
- ***Decision making techniques***
- ***Time series Analysis***
- ***GLM Analysis***
- ***Multivariate Analysis***
- ***Data Analysis in Multi-agent systems***

Question 1

In a wireless network, four sensors sense and analyze their own datasets.

(a) Model each sensor e.g. $S_j, j = 1, \ldots, 4$ with $N(j, 25)$, and generate 100 samples for each sensor. Frame all samples into one dataset.

(10 Marks)

(b) Provide the descriptive analysis for your dataset (e.g. summary, boxplot, …). Interpret your quick insight about the simulated dataset.

(5 Marks)

(c) Make a decision whether the population variance of the first sensor (σ_1^2), is significantly different from the mean of the second sensor (σ_2^2) at the level $\alpha = 0.05$. To do so,

I. List the assumptions.

(5 Marks)

II. State the null and alternative hypotheses.

(5 Marks)

III. What is your decision rule and explain your decision?

(5 Marks)

IV. Provide the 95% confidence interval for the ratio of the variance.

(5 Marks)

(TOTAL: 35 Marks)

Question 2

Use dataset available on http://www.stat.ufl.edu/~winner/data/clotthes_expend.csv , apply time series analysis, consider **sales.b** as your time series variable:

(a) Validate the assumptions using graphical visualization.

(5 Marks)

(b) Fit the optimized model for **sales.b** and provide the coefficient estimates for the fitted model.

(5 Marks)

(c) What is the estimated order for AR and MA?

(5 Marks)

(d) Forecast h=10 step ahead prediction of **sales.b** on the plot of the original time series.

(5 Marks)

(Total: 20 Marks)

Question 3

Use dataset available on http://www.stat.ufl.edu/~winner/data/HVAC_perform.csv,

(a) Suggest an appropriate GLM to model **powerp** to other numerical variables.

(5 Marks)

(b) Specify the significant variables on **powerp** at the level of α=0.05, and estimate the parameters of your model. **(5 Marks)**

(c) Predict the value of **powerp** for an optional choice.

(5 Marks)

(d) Provide predictions with their confidence interval.

(5 Marks)

(Total: 20 Marks)

Question 4

Use dataset available on
http://www.stat.ufl.edu/~winner/data/iran_rock.csv,

(a) Perform ANOVA.

(10 Marks)

(b) Load the dataset available on
http://www.stat.ufl.edu/~winner/data/esp_studies1.csv,
Apply PCA, and identify the important principle components involving at least 80% of dataset variation. Explain your decision strategy.

(10 Marks)

(Total: 20 Marks)

Question 5

Use the simulated dataset in Question 1 in order to

(a) Adopt a centralized scheme to sensor 1 and 2 and sketch the graphical scheme.

(10 Marks)

(b) Compute the normalized weights and find the global arithmetic mean. Please compute the global solution using R.

(15 Marks)

(Total: 25 Marks)