Dublin Business School
excellence through learning

QQI

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

RESIT EXAMINATIONS

Module Code:	**B8IT109**
Module Description:	**Advanced Data Analytics**
Examiner:	**Dr Shahram Azizi Sazi**
Internal Moderator:	**Mr Paul Laird**
External Examiner:	

Date: 20th December 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 .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 traffic management system (TMS), two traffic sensors are designed. The first sensor records the number of cars (X) per minute and the second sensor measures the waiting time of each car (T).

(a) What is the appropriate probability model of X ? (based on the pervious information, the number of cars per minute is 7 on average) **(5 Marks)**

(b) What is your proposed probability model of T? (choose optionally the value of the parameter for your proposed model). **(5 Marks)**

(c) Generate 100 samples for each sensor. Frame all samples into one dataset.

(5 Marks)

(d) Provide the descriptive analysis for your dataset (e.g. summary, boxplot, …). Briefly document your insight about the simulated dataset.

(5 Marks)

(e) In a multi-agent system (MAS), four agents are active and each agent e.g. A_j, j=1,…,4 is normally distributed with N(j,16). Simulate 40 samples for the agents one and four and test whether the population mean of the first agent (μ_1), is significantly different from the mean of the fourth agent (μ_4) 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 $(\mu_1 - \mu_2)$.

(5 Marks)

(TOTAL: 40 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://users.stat.ufl.edu/~winner/data/nfl2008_fga.csv

(a) Use LDA to classify the dataset into few classes so that at least 90% of information of dataset is explained through new classification. (**Hint**: model the variable "**qtr"** to variables "**togo**", "**kicker**", and "**ydline**"). How many LDs do you choose? Explain the reason.

(5 Marks)

(b) Apply PCA, and identify the important principle components involving at least 90% of dataset variation. Explain your decision strategy? Plot principle components versus their variance (**Hint**: to sketch the plot use the Scree plot).

(5 Marks)

(c) Split the dataset into two sets of variables so that **X**=(togo,kicker,ydline) and **Y**=(distance, homekick). Apply canonical correlation analysis to find the cross-correlation between **X** and **Y**. What is the correlation between ***ydline*** and ***distance***? **(5 Marks)**

(d) Use K-means clustering analysis to identify the most important classes. How many classes do you select? Why?

(5 Marks)

(Total: 20 Marks)

Question 5

Use the simulated dataset in Question 1 in order to

(a) Adopt a distributed scheme so that A1 reports the knowledge (dataset) to A2, and A3 reports the knowledge to A4 and sketch the graphical scheme.

(10 Marks)

(b) Compute the normalized weights and find the arithmetic mean for the agent A2. Find the geometric mean for the agent A4.

(15 Marks)

(Total: 25 Marks)