ROBERT GORDON UNIVERSITY ABERDEEN

Assessment Brief Template

Academic Year	**2023/24**
Semester	**2**
Module Number	**CMM 703**
Module Title	**Data Analysis**
Assessment Method	**Coursework**
Deadline (time and date)	**22/04/2023**
Submission	**Assessment Dropbox in the Module Study Area in CampusMoodle.**
Word Limit (see Assessment Word Limit Statement)	**Not more than 2000 words, including codes.**
Module Co-ordinator	**Sameera Viswakula**

What knowledge and/or skills will I develop by undertaking the assessment?
Develop in-depth knowledge of the data analytic lifecycle and specialised programming knowledge through the statistical programming language R.
On successful completion of the assessment students will be able to achieve the following Learning Outcomes: 1. Critically appraise data transformation methods for statistical analysis. 2. Justify analysis methods and conclusions by selective and critical use of relevant theories. 3. Design, implement and evaluate the data analytic lifecycle stages: clean, transform, analyse and visualise. 4. Communicate conclusions, insights and recommendations to a wider audience by tailoring them at different levels of detail.
Please also refer to the Module Descriptor, available from the module Moodle study area.

What is expected of me in this assessment?

Task(s) – content

This coursework aims to extend the R programming and data analysis implementation from lectures. The secondary aim is to test your ability to apply and transfer your knowledge to a real-life scenario.

This contributes 4 subgrades towards the final grade. Please refer to the feedback grading for the details.

Task 1

Consider the "SriLanka_Weather_Dataset.csv" file. It consists of a comprehensive collection of weather data for 30 prominent cities in Sri Lanka, covering the period from January 1, 2010, to January 1, 2023. Generate two important plots that can be used to visualize important aspects of the dataset. Discuss how you can improve the plots. [LO1, LO2, LO3]

Task 2

Consider the "lepto_data.csv" dataset. The descriptions of the variables are given in the "lepto_description.xls" file. This dataset contains demographic and clinical data of 1735 patients related to leptospirosis. The variable "Final" (last column) reports the leptospirosis status of the patient (1-confirmed, 2-not detected). [LO1, LO2, LO2, LO4]

Task 2.1

Do a thorough descriptive analysis and identify the patterns and potential significant variables. Use appropriate plots and tables.

Task 2.2

Fit a suitable predictive model to predict the leptospirosis status of the patient using a proper train dataset. [You may use transformations, etc. techniques to improve the model]

Task 2.3

Get the predictions from the model for the corresponding test dataset.

Task 2.4

Now fit a suitable predictive model taking only the non-clinical variables and get the

What is expected of me in this assessment?

predictions for the same training and testing datasets. Compare your prediction metrics and discuss the answers.

Task 3

Write an R function to do the following tasks. When a dataset is fed to the function, your function should: [LO3, LO4]

Task 3.1

Identify qualitative and quantitative variables in the dataset.

Task 3.2

Count the missing values in each variable. Impute the missing values using: the mean value of the variable if it is numeric; the mode of the variable if it is categorical.

Task 3.3

Identify univariate outliers for each numeric variable.

Task 3.4

Summarize each variable using a proper visualization tool for the respective variable (eg: histogram, boxplot etc.).

Task 3.5

When the response variable is specified as an argument, it should run the best predictive model for that response category (consider only continuous and binary response variables) and select features considering all other meaningful variables. Your function should print relevant diagnostic metrics and plots for the selected model.

Task 3.6

Implement the above functions in an R Shiny app/dashboard.

If you do not attend for the viva, your grade will downgrade to an F.

Task(s) – format

You will be expected to provide R codes and interpretations/answers as a single PDF report generated by the RStudio (File menu -> Compile Report -> as PDF). Please label the answers properly on your document and adhere to the word limit provided. Your PDF (single file) file should be uploaded to the link provided on the campusmoodle by the deadline provided. If you fail to show up for a viva after submission, you will be given an F grade.

How will I be graded?	
A grade will be provided for each criterion on the feedback grid which is specific to the assessment. The overall grade for the assessment will be calculated using the algorithm below.	
A	At least 80% of the feedback grid to be at Grade A, and normally 100% of the feedback grid to be at Grade C or better.
B	At least 80% of the feedback grid to be at Grade B or better, and normally 100% of the feedback grid to be at Grade D or better.
C	At least 80% of the feedback grid to be at Grade C or better, and at least 80% of the feedback grid to be at Grade D or better.
D	At least 80% of the feedback grid to be at Grade D or better, and at least 80% of the feedback grid to be at Grade E or better.
E	At least 50% of the feedback grid to be at Grade E or better.
F	Absence of viva or failing to achieve at least 50% of the feedback grid to be at Grade E or better.
NS	Non-submission.

ROBERT GORDON UNIVERSITY ABERDEEN

Feedback grid.

GRADE	A	B	C	D	E	F
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT** Outstanding Performance	**COMMENDABLE/VERY GOOD** Meritorious Performance	**GOOD** Highly Competent Performance	**SATISFACTORY** Competent Performance	**BORDERLINE FAIL**	**UNSATISFACTORY** Fail
Task 1 **(x %)** **Weight:** 1	Excellent use of graphs to convey the information in the data using R. Accurate use of colours, axis naming and labelling legends. Excellent discussion of expected and unexpected results.	Very good use of graphs to convey the information in the data using R. Accurate use of colours, axis naming and labelling legends. Very good insightful discussions of results.	Good use of graphs to convey the information in the data using R. Good use of colours, axis naming and labelling legends. Good insightful discussions of results.	Satisfactory use of graphs to convey the information in the data using R. Satisfactory use of colours, axis naming and labelling legends. Some insightful discussions of results.	Some evidence of the use of graphs to convey the information in the data using R. Satisfactory use of colours, axis naming and labelling legends. Some insightful discussions of results.	Lack of evidence of the use of graphs to convey the information in the data using R. Fail to use proper colours, axis naming and labelling legends. Important insightful discussions of results are missing
Task 2 **(x %)** **Weight:** 2	Complete implementation of two classification models. Excellent comparison of the two models using proper metrics. Excellent use of proper variable selection methods and come up with a parsimonious model.	Very good implementation of two classification models. Very good comparison of the two models using proper metrics. Very good use of proper variable selection methods and come up with a parsimonious model.	Good implementation of two classification models. Good comparison of the two models using proper metrics. Good use of proper variable selection methods and come up with a decent model.	Partial implementation of two classification models or use of incorrect models. Use of Some comparison of the two models using proper metrics. No use of proper variable selection methods.	Implementation of at least one classification model. No of proper variable selection methods.	No implementation of at least one classification model or use of wrong models. No comparison of models. No of proper variable selection methods.
Task 3 **(x %)** **Weight:** 2	Excellent implementation of all sub functions as a single function and develop an excellent R Shiny dashboard.	A very good implementation of all sub functions as a single function and develop a very good R Shiny dashboard.	A good implementation of all sub functions as a single function and partially develop an R Shiny dashboard.	Some implementations of sub functions and partially develop an R Shiny dashboard.	Some implementations of sub functions and not developing an R Shiny dashboard.	Incorrect development or no development of at least one sub functions.
Viva **(x %)** **Weight:** 1	Excellent knowledge of important data analysis techniques and R functionality.	A very good knowledge of important data analysis techniques and R functionality.	A good knowledge of important data analysis techniques and R functionality.	Some knowledge of important data analysis techniques and R functionality.	Lack of knowledge of important data analysis techniques and R functionality.	Lack of knowledge of most of the data analysis techniques required for the CW.

Coursework received late, will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost.

What else is important to my assessment?

What is plagiarism?

"Plagiarism is the practice of presenting the thoughts, writings or other output of another or others as original, without acknowledgement of their source(s) at the point of their use in the student's work. All materials including text, data, diagrams or other illustrations used to support a piece of work, whether from a printed publication or from electronic media, should be appropriately identified and referenced and should not normally be copied directly unless as an acknowledged quotation. Text, opinions or ideas translated into the words of the individual student should in all cases acknowledge the original source" (RGU 2022).

What is collusion?

"Collusion is defined as two or more people working together with the intention of deceiving another. Within the academic environment this can occur when students work with others on an assignment, or part of an assignment, that is intended to be completed separately" (RGU 2022).

For further information please see Academic Integrity.

What is the Assessment Word Limit Statement?

It is important that you adhere to the Word Limit specified above. The Assessment Word Limit Statement lists what is included and excluded from the word count, along with the penalty for exceeding the upper limit.

What if I'm unable to submit?

- The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so.
- If you require an extension, you should complete and submit a Coursework Extension Form. This form is available on the RGU Student and Applicant Forms page.
- Further support is available from your Course Leader.

What else is important to my assessment?

What additional support is available?

- RGU Study Skills provide advice and guidance on academic writing, study skills, maths and statistics and basic IT.
- RGU Library guidance on referencing and citing.
- The Inclusion Centre: Disability & Dyslexia.
- Your Module Coordinator, Course Leader and designated Personal Tutor can also provide support.

What are the University rules on assessment?

The University Regulation 'A4: Assessment and Recommendations of Assessment Boards' sets out important information about assessment and how it is conducted across the University.