Department of Computer Science

CS5801 Quantitative Data Analysis

Assessment/Coursework for 2024/25

TABLE OF CONTENTS

Assessment Title	Quantitative Data Analysis
Module Leader	Dr Isabel Sassoon, Prof Martin Shepperd
Distribution Date	30/10/2024
Submission Deadline	6th of January 2025 11.00 am UK time
Feedback by	27/02/2025
Contribution to overall module assessment	100 %
Indicative student time working on assessment	70 Hours
Word or Page Limit (if applicable)	2000 Words (not including code and references)
Assessment Type (individual or group)	Individual Coursework

MAIN OBJECTIVE OF THE ASSESSMENT

The coursework is based on undertaking an authentic analysis of a real-world data set. It is a shared assessment block for CS5701 Quantitative Data Analysis and CS5702 Modern Data.

This assessment offers an opportunity to bring together your skills from CS5701 (Quantitative Data Analysis) and CS5702 (Modern Data). Note that for students on MSc Data Science and Analytics and MSc. Artificial Intelligence there is also a second shared assessment block CS5802 which takes the form of a written closed book examination.

DESCRIPTION OF THE ASSESSMENT

In order to work on the assessment, you will be provided with:

(i) A data set (student_data.Rda)

(ii) The data set metadata is provided at the end of this assessment brief, (note students will generate a subset of the dataset by following the instructions provided in the proforma Rmd and explained below)

(iii) An submission R markdown proforma with questions to guide your analysis (proforma-submission-24-25.Rmd). **This submission rmd file is the only file that will be used for marking,**

(iv) A supporting R markdown proforma to use to contain additional material (proforma-supporting-24-25.Rmd). **This file will also need to be submitted**.

In order to submit the assessment you will need to submit two files:

(i) A proforma-submission-24-25.Rmd file. This is the only file that will be used for marking and needs to be clearly named (see submission instructions on page 5 for details)

(ii) A proforma-supporting-24-25.Rmd file. This file will also need to be submitted. (see submission instructions on page 5 for details)

In order to answer the questions in the submission R markdown you will need to explore different options so use the supporting R markdown document to contain additional parts of the exploration code and outputs. For

example: you may do a more comprehensive EDA in order to decide which three elements you find most informative.

Generating your personal data sets

1. Each student should use a subset from the overall dataset student_data.Rda which can be downloaded from the Brightspace page for CS5801.
2. The subset of the data that you will work on depends on your student id.
3. At the top of the proforma-submission-24-25.Rmd file provided (can be downloaded from the Brightspace page of CS5801) there are instructions on how to use your student id to obtain the subset of data you need to work on. The same should be done in proforma-supporting-24-25.Rmd.
4. The code for sub setting is embedded in the RMarkdown template. You only need to configure it for your student id.
5. If you are uncertain, please check!

General Guidance:

1. You are expected to use R in a RMarkdown file for your analysis.
2. Use the proforma-supporting-24-25.Rmd and the questions to work on your full analysis then for submission purposes you need to complete all the sections in the proforma-submission-24-25.Rmd.
3. Update the YAML to include your name and other identifier information.
4. Follow the principles of 'literate programming' so choose meaningful variable and function names and add comments.
5. You need to submit the supporting file but the marking will be done on the contents of proforma-submission-24-25.Rmd. Make sure that if you claim to have tried an approach in the submission file then the relevant code is in the proforma-supporting-24-25.Rmd.
6. Where appropriate cite external sources and add a bibliography at the end of your main report.
7. The .Rmd file should be professionally presented with good structure, an absence of spelling errors and other typos and written in an appropriate style (i.e., simple to the point, unemotive language).
8. Make sure you respect the 2000 word limit as well as the specific section word limits as we discourage excessive padding, so unnecessary words and waffle will militate against professional presentation. This word count does not include contents of code chunks or references.
9. Where answers to sections contain more words than those in the word limit for that section (See the marking criteria below) then words beyond the maximum word count (with a discretionary 20% buffer) will be disregarded and will mean you lose marks.
10. Avoid including generic definitions and focus on contextualising your explanations and findings from your analysis of your specific data set subset.
11. Sometimes even suitable models do not have good fit due to the nature of the data. In such circumstances you will not be penalised.
12. Whilst we encourage collaboration and sharing of ideas this is an individual report and so must be based on your own understanding, analysis and words. WiseFlow automatically cross-compares all submissions.
13. WiseFlow also has a plagiarism detector for external sources. We encourage you to use such sources including R packages, code, ideas for data analysis and other statistical sources, but you must acknowledge the sources. In other words, do *not* attempt to pass off the work of others as your own.
14. If you have questions, please post a question on the Brightspace CS5801 Discussion Forum or ask one of us. Don't guess!

LEARNING OUTCOMES AND MARKING CRITERIA

LO1: Design and implement methods and protocols for data preparation and exploration using advanced statistical techniques.
LO2: Apply these methods on real data to generate novel insight, critically evaluate its value and design a framework for data management and sharing.

Below we give the details of the marking scheme.
(Note: the mark scheme below is applied to contents the submission Rmd file. This submission rmd file is the only file that will be used for marking)

section	question no.	question	Guidance	Marks	word count	Mapping to LO
0		Presentation, clarity and good practice for the report	Your submission Rmd file should be clearly and professionally presented with appropriate use of cited external sources You report should respect the word counts in each section and a supporting proforma.rmd should also be submitted. The report should not include generic definitions or statements that are not linked to your analysis	10		LO2
0		Programming	The code needs to run end to end in both the submission Rmd (this one) and the supporting one. It should also be easy to understand, with well-documented code following the principles of literate programming. The code documentation needs to be specific to your analysis (The supporting rmd document also needs to be submitted)	10		LO1
1		Organise and clean the data				
	1.1	Subset the data into the specific dataset allocated	Use R code to correctly select the subset of data allocated.	5		LO1
	1.2	Data quality plan	Provide a description of the steps you took to assess the quality of this data. Include and refer to all variables/columns from the data set.	5	200	LO2
	1.3	Data quality findings	List three data quality issues you have identified. For each one include the code that supports the finding in the code chunks provided.	9	300	LO1
	1.4	Data Cleaning	For the three issues listed in 1.3 explain how you addressed each of the issues. Include justification for each of the approaches to address the issue, alternative approaches considered and supply the code within the code chunks provided	6	200	LO1
2		EDA				

	2.1	EDA Summary	In this section include: (i) A brief explanation of your approach to exploring the data set also with respect to the dependent variables. (ii) three elements (graphical or numerical/textual/tabular) of your exploratory analysis that you find are most informative. Include the code for each one (in the separate chunks provided) and explain and interpret the output and also articulate why this is informative and helpful. (You should provide the code for you exploration in the supporting proforma.)	15	300	LO2
3		Modelling dependent variable 1				
	3.1	Explain your analysis	The aim of the analysis is to model the variable Grade Point Average (variable name: grade_point_average). Describe and justify the steps you took to model this variable (don't include or repeat the data cleaning and EDA plan). Explain: What methods you used and why? How were the findings from EDA incorporated? What model selection approach did you take? How did you address weaknesses in models? What alternative approaches did you consider? You should be specific to your data set in your justifications Do not include or repeat the data cleaning and EDA plan Do not include code for all the models you considered (this should be in the supporting proforma)	10	400	LO1
	3.2	Provide a model for grade point average	Explain, interpret and justify from all the models attempted the model you think is best. Include the code for that model in a code chunks alongside any code related to diagnostics.	10	100	LO2
4		Modelling dependent variable 2				

	4.1	Model the likelihood of completing an extended project	Model the likelihood of completing an extended project (using the completed.extended.project variable provided) The aim of the analysis is to model whether a student has completed an extended project or not. (i.e., involving the binary target attribute). *Describe and justify the steps you took to model whether a student completed an extended project or not (don't include or repeat the data cleaning and EDA plan or don't include the code for all the models attempted). Do explain: What methods you used and why? How were the findings from EDA incorporated? What model selection approach did you take? How did you address weaknesses in models? What alternative approaches did you consider?	10	400	LO1
	4.2	Propose one model	Justify and propose one model. Describe, interpret and critique it. Include the code for this proposed model in a code chunk provided	10	100	LO2

SUBMISSION INSTRUCTIONS

You must submit your coursework as **two .rmd files** on WISEflow by 06/01/2025 at 11am. You can follow the link to WISEflow through the module's section on Brightspace or login in directly at https://uk.wiseflow.net/brunel. The name of your files should follow the normal convention and must therefore include your student ID number and indicate whether it is submission or supporting file (e.g., 0612345-submission-24-25.Rmd and 0612345-supporting-24-25.Rmd). It can also include the module code (e.g., CS5801_0612345-submission-24-25.rmd and CS5801_0612345-supporting-24-25.rmd).

AVOIDING ACADEMIC MISCONDUCT

Before working on and then submitting your coursework, please ensure that you understand the meaning of plagiarism, collusion, and cheating (including contract cheating) and the seriousness of these offences. Academic misconduct is serious and being found guilty of it results in penalties that can reduce the class of your degree and may lead to you being expelled from the University. Information on what constitutes academic misconduct and the potential consequences for students can be found in Senate Regulation 6.

You may also find it useful to read this PowerPoint presentation which explains, in plain English, the different kinds of misconduct, how to avoid (even accidently) committing them, how we detect misconduct, and the common reasons that students give for engaging in such activities.

If you are experiencing difficulties with any part of your studies, remember there is always help available:

- Speak to your personal tutor. If you're not sure who your tutor is, please ask the Taught Programmes Office (TPOcomputerscience@brunel.ac.uk).
- Alternatively, if you prefer to speak to someone outside of the Department you can contact the Student Support and Welfare team.

Expectations of Artificial Intelligence Use

The University has general guidance on using artificial intelligence in your studies.

Generative AI (GenAI) and Large language models (LLMs) such as chatGPT and Claude all have the potential to answer questions on code and material relevant to this assessment. Using AI-generated content and presenting it as your original work in this assessment is strictly prohibited. Despite its many useful services, such as, information search and retrieval, concept explanations and proofreading, it is crucial to recognize its inherent limitations. For instance, responses generated by AI can be overly general, inaccurate, biased, or even fabricated. They often lack proper references and detailed insights and may pose challenges in terms of intellectual property rights and data privacy.

If you make use of Generative AI tools in any part of this assessment, it is essential to:

- Acknowledge that you have used GenAI, how you used it and reference any of the AI contents used in your submission (including code). If you need to reference or cite content from an AI tool, use the name of the tool used (e.g. Claude).
- Avoid copying explanations, text, or code directly from a GenAI tool into your assessment submission. Use your own words to answer the questions asked in the assessment. If you use GenAI , to help with code rigorously test and alter the code to fulfill the requirements of the assessment, also comment the code in your own words.
- Understand any code you submit or explanation you provide and be ready to explain it verbally. In the event of any concerns regarding the integrity of your work, an oral examination may be scheduled for further evaluation.

Misuse of Generative AI can lead to academic misconduct for more details see the link to the general guidance from the University above.

Late Coursework

The clear expectation is that you will submit your coursework by the submission deadline stated in the study guide. In line with the University's policy on the late submission of coursework (revised in July 2016), coursework submitted up to 48 hours late will be accepted but capped at a threshold pass (D- for undergraduate or C- for postgraduate). Work submitted over 48 hours after the stated deadline will automatically be given a fail grade (F).

Please refer to the Computer Science student information pages and the Coursework Submission Procedure pages for information on submitting late work, penalties applied and procedures in the case of Extenuating circumstances.

APPENDIX – METADATA

This data set includes data related to one large cohort of secondary school students from different schools.

Variable Name	Description
student_id	Student ID
entry_exam_mark	Students entering the school have taken an exam. Marks are between 0 and 100.
sat_score	Students have taken the SAT test that is graded between 0 and 1600.
percentage_absence	Student's absence is computed as a percentage
free_school_meals	This column is Yes if the student is entitled to financial support covering the cost of their school lunch. No otherwise.
grade_point_average	This is the grade students achieved at graduation
completed.extended.project	This is Yes when students completed an independent project during the school year. No Otherwise.
month_of_birth	Month of Birth of students
commute_method	This shows the method of travel to and from school for each student. (Bicycle, Bus, Car, Walking, Other)