JAMES COOK UNIVERSITY
Faculty of Science Engineering and Information Technology
School of Maths, Physics and IT

CAPSTONE PROJECT – FOUNDATIONS OF DATA SCIENCE

OVERVIEW

This assessment involves writing a report that summarises a data science related investigation that you have conducted on data that you have collected yourself. The investigation must involve the main topics covered in the subject, most noticeably data pre-processing (representation, wrangling, tidying) and exploratory data visualisation **using R/RStudio**.

It is a merger of Assessments 3 (Exploratory Visualisation) and 4 (Pre-Processing – Parts A and B), however neither the dataset nor the pre-processing/exploratory steps to be carried out will be provided, you have to make independent choices and decisions.

You will need to find your own data using good practices. Your dataset **cannot** be smaller than 1000 observations of 5 variables, except if the targeted data science problem to be addressed relates to spatial-temporal data, case in which less than 5 dimensions could be allowed.

Preferably, you should use a dataset relevant to your place of work. **Do not** use data from textbooks or from R packages. **Do not** use data from the same public sources that have been used in the subject (e.g. UCI repository). You can use public data, but the data should be appropriate for addressing a relevant data science problem.

You don't need to solve this entire data science problem in your investigation, but you need to clearly indicate what the targeted problem would be about and how your project can contribute towards addressing it.

You have to write a report with details about the problem in question, the data, the methods, results, analyses and findings. You might like to look online for research papers for examples of how to shape your report. Obviously many of these papers will have undergone extensive work to collect their data, we don't expect that for you.

We also don't expect you to win a Nobel prize with this assessment. Ideally, you will be able to demonstrate that: (a) you have grasped important concepts associated with this subject, most noticeably data pre-processing and exploratory visualisation; and (b) you can communicate your investigation in a formal written manner.

Regarding (a), we expect that your investigation will include at least six (60% or more) of the following topics:

1. Data representation
2. Unstructured to Structured data
3. Data cleaning
4. Type conversion
5. Missing value imputation
6. Gathering/Spreading
7. Data subset selection and/or subsampling
8. Group-based data summarisation
9. Variable selection and/or transformation
10. Exploratory visualisation using ggplot2

Regarding (b), the **main body** of the report (containing title, abstract, introduction, data, methods, results and discussion, and conclusions) **cannot exceed 5 (five) A4 pages in 12pt Roman style font using single line spacing**. A **maximum of 5 (five) additional pages** are allowed for bibliographic references and appendices with any supporting material that you may want to include (e.g. your R codes). Therefore, your report **cannot exceed ten (10) pages in total**. Only the main body and references will be formally assessed for grading, though the additional material can help clarify any issues that may arise during the marking process. Further details about the report structure are provided in the following section.

REPORT STRUCTURE

The report should have the following sections marked clearly:

- **Title:** In today's busy world, it is very important to make the most of your title. Make the title 'eye-catching', informative and an accurate representation of the contents of the report.
- **Abstract**: The abstract provides a short sharp overview of the contents in the report and will be around 200 – 300 words. The abstract has five parts:
    i. Introductory statement: background to the study, important issue(s) the report addresses. (approximately 1 to 2 sentences)
    ii. Purpose of the report: state the objectives (1-2 sentences)
    iii. Methodological approach: overview the data and methods (2-3 sentences)
    iv. Findings or Achievements: list one or two of the main findings or achievements from your investigation (1-2 sentences)
    v. Conclusions and Implications: what conclusions can be drawn from your investigation? How can the findings/achievements in your report deliver a benefit to people, things, systems or processes? (1-2 sentences)
- **Introduction**: The introduction sets the scene for the investigative efforts. It provides motivation for the work and relevant background information and references that will enable the reader to put in context the key objectives and achievements in your report. Address the important issues that have motivated your investigation. At the end of the introduction clearly state the objectives of the report. Do not put any results from your investigation in the introduction. Do not discuss details about the data and methods in this section. Do not discuss your conclusions or key findings in the introduction.
- **Data**: This section should provide details about how the data was obtained and what the data represent. You should include information such as:
    i. What the source of the data is.
    ii. How the data was originally collected (e.g. from an experiment or observational study).
    iii. The sample size.
    iv. The number and types of variables.
    v. Any known interventions or pre-processing that precede the ones described in your report.
    vi. Any other information that is relevant to the understanding and assessment of your work/report.
- **Methods**: This section should summarise the data science methods that were used to process and to analyse the data, as well as the software version used to generate the results. To cite R-Studio type *RStudio.Version()* from the command line. The methods should be appropriate to ensure that the objectives of the paper are met. At times, it may be helpful to interleave your text with a description of key calls to R functions that generated relevant results that you may want to highlight. E.g. "The **lm** command with default settings for the arguments was used to produce a simple linear regression model between y and x in R-Studio". It is important to provide the sufficient level of details so that your methodology could be repeated by an independent person, while being clearly and objectively presented so that it can be understood without the need to check your complete R code.
- **Results and Discussion**: This section presents and discusses the results. The discussion centres on the outputs from the pre-processing and exploratory visualisations that you have performed. For example, what are the main outcomes? Why are they useful and what for? How are they interesting and why? Etc. In particular, how do the results align with the goals set in the introduction? What are the main achievements and their implications?

- **Conclusions**: Final remarks about the key achievements of the investigations and what makes them "interesting" or "useful", right now or for future work. Achievements or findings should be contrasted with the original objectives or hypotheses of the project. Make sure that you mention any limitations of your work here. Limit the conclusions to no more than two or three paragraphs.
- **References**. List the sources your investigation has drawn from. Note that all references should be referred to in the text.
- **Appendices (optional)**: Add any supporting materials (possibly your detailed R codes) that might be useful to help assess your work.

FORMAT

The **main body** of the report must be presented in 12pt Roman style font on no more than **5 (five) A4 pages**, using single line spacing. Either a single column or double column format may be used.

References and appendices can be listed on **at most 5 (five) additional pages**.

In total, **the report cannot exceed 10 pages**.

WARNING: **only the main body and the references will be formally assessed and graded**.

IMPORTANT NOTES

1. The **entire project** must be accomplished using **R/RStudio**. Any calculations, visualisations, results, etc. produced using software other than R/RStudio (e.g. Excel, Tableau, etc.) is **not** accepted and therefore will not be assessed. Exploratory visualisation must use package **ggplot2**, rather than functions from base R or other R packages. The report itself can be written using a text editor of your choice (e.g. Microsoft Word or alike); R Markdown is also accepted, but it is not compulsory.
2. If you opt to not submit your R codes appended to the report, the instructor and facilitators reserve the right to ask you to do so if more details or evidence are deemed required to properly assess your work. Refusal to comply with this requirement may incur in your work being considered as not delivered.

A WORD ON PLAGIARISM AND SELF-PLAGIARISM:

Plagiarism is the act of using another's words, works or ideas from any source as one's own. Plagiarism has no place in a University. Student work containing plagiarised material will be subject to formal university processes.

In case significant portions of your own previous work (e.g. a report for a related subject you did in this or any other university) is recycled in a way that it could be fully or partially graded twice ("double-dipping"), this is considered **self-plagiarism** and will not be tolerated.

JAMES COOK UNIVERSITY

Faculty of Science Engineering and Information Technology

School of Maths, Physics and IT

MARKING SCHEME

Please adhere to the strict formatting requirements. The report will not be assessed if it is not formatted appropriately.

Total marks possible 120.

Dimension	**Sophisticated [100% marks]**	**Competent [50% marks]**	**Needs Work [0% marks]**
Title [2 marks]	The title is a concise (less than 20 words) and accurate reflection of the contents of the report. Author is listed below the title.	The title is a concise (less than 20 words) and moderately reflects the contents of the report. Author is listed.	The title is not informative or exceeds the word length or Author not listed.
Abstract [6 marks]	Clearly addresses the five parts of the abstract so that the reader has a clear overview of the reports.	Partially addresses the five parts of the abstract and or addresses all five parts but the writing is not clear in places.	Unclear, does not overview the report, or the writing is poor overall and mostly unclear
Introduction [16 marks]	Position and exceptions, if any, are clearly stated. Organization of the argument is completely and clearly outlined and implemented.	Position is clearly stated. Organization of argument is clear in parts or only partially described and mostly implemented.	Position is vague. Organization of argument is missing, vague, or not consistently maintained.
Data [20 marks]	Data are suitable, the report explains how the data were obtained, and all of the following information items (whenever applicable) are clearly explained: i. What the source of the data is. ii. How the data was originally collected (e.g. from an experiment or observational study). iii. The sample size. iv. The number and types of variables.	Data are suitable, the report explains how the data were obtained, and most of the applicable data information items are addressed and reasonably explained.	Little information/explanation about the data is provided and/or the grammar structure is difficult to follow and/or the data do not meet the minimum requirements.

	v. Any known interventions or pre-processing that precede the ones described in your report. vi. Any other information that is relevant to the understanding and assessment of your work/report.		
Methods [28 marks]	Lists all the steps in order in which they were performed to pre-process and/or explore the data. These steps, if executed appropriately and interpreted appropriately, will ensure that the objectives of the report are clearly met. At least 6 of the following targeted key topics from the subject have been explored and explained in depth: 1. Data representation 2. Unstructured to Structured data 3. Data cleaning 4. Type conversion 5. Missing value imputation 6. Gathering/Spreading 7. Data subset selection and/or subsampling 8. Group-based data summarisation 9. Variable selection and/or transformation 10. Exploratory visualisation using ggplot2	Most of the steps are listed and explained, but some details are a little hazy or questionable. At least 4 of the targeted key topics from the subject (listed in the leftmost column) have been reasonably explored and explained.	The methods clearly will not allow the objectives of the report to be met and/or the details of methodological steps and procedures are very difficult to follow and/or the listed key topics from the subject have been poorly or not appropriately explored.
Results and Discussion [22 marks]	The results and discussion are explained correctly, clearly, and in sufficient detail. The results and discussion clearly follow from the data collection and the methods.	The results and discussion are explained correctly, clearly and in sufficient detail most of the time. There exists a connection of some type between the results/discussion and the data collection and methods.	One or more of the items discussed in the middle column are missing.

Conclusion [10 marks]	The original objectives and/or hypotheses are restated and contrasted against the obtained achievements and/or findings. The conclusion summarizes and draws a clear, effective conclusion of the investigation and enhances the impact of the report – e.g., it provides a recommendation or action that should be undertaken in the future. It may also highlight unavoidable limitations of the investigation.	Conclusion is clearly stated and connections to the original objectives and/or hypotheses are mostly clear.	Conclusion may not be clear and/or the connections to the work reported are incorrect or unclear or just a repetition of the findings without a suitable summarisation and interpretation and/or the underlying logic has major flaws.
Writing [16 marks]	Report is coherently organized and the logic is easy to follow. There are no spelling or grammatical errors and terminology is clearly defined. Writing is clear and concise and persuasive. Each Figure/Table will be numbered, followed by a caption, and referred to in the body of the text, most noticeably in the results and/or discussion section. The Figures/Tables provided reinforce the most relevant achievements of the work. All references have been listed and referred to in the appropriate places in the body of the text and listed at the end of the report. At least 4 references have been provided.	Report is generally well organized and most of the argument is easy to follow. There are only a few minor spelling or grammatical errors, or terms are not clearly defined. Writing is mostly clear but may lack conciseness. Each Figure/Table will be numbered, followed by a caption, and referred to in the body of the text, most noticeably in the results and/or discussion section. Most references have been listed and referred to in the appropriate places in the body of the text and listed at the end of the report. At least 4 references have been provided.	Report is poorly organized and difficult to read – does not flow logically from one part to another. There are several spelling and/or grammatical errors; technical terms may not be defined or are poorly defined; figures/tables and/or references are sloppy or missing. Writing lacks clarity and conciseness.