|  |
| --- |
| OVERVIEW For the continuous assessment you are required to conduct and report on a statistical analysis to investigate a question for a given dataset. The dataset is available for download from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/student+performance>) where you will find a description. It is also used in the following paper which also provides a dataset descriptor:  P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. (<https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf>)  Please ensure that you include this citation in the report you submit.  For the purposes of the CA for this module you should consider that this is training data only. The dataset does not contain data from sufficient years to be able to fully fit a model, but it does contain enough for use to build and assess the fit of an initial model.  For this part of the assignment you are required to identify a number of concepts for which variables are included in the data (or for which you can derive measures from variables in the dataset), inspect the relevant variables and report your findings including relevant descriptive statistics and visuals and present the outcomes of preliminary exploratory analysis of correlation and difference. This part of the assignment is worth 50% of the CA for module as marked out of 100%. |

|  |
| --- |
| NOTES  1. Unfair practice is a very serious offence in TU Dublin and you must acknowledge any material used by including a referenced bibliography in your report. Any issues will be investigated and those considered serious will be handled via the TU Dublin Plagiarism policy (details are available in the General Assessment Regulations). 2. You are required to treat the dataset provided ethically and conduct your statistical analysis ethically. As such you should adopt the guidelines for ethical statistical practice provided by the American Statistical Association <https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx> 3. You are required to adopt the APA guidelines for reporting statistics and for citation <https://apastyle.apa.org/> and report your tests adhering to APA conventions (this style guide should provide you with the information you need <http://spss.allenandunwin.com.s3-website-ap-southeast-2.amazonaws.com/Files/APAStyle.pdf>) 4. Assignments must be submitted via Brightspace through the assignment section. Email submissions will be ignored. 5. Extensions due to acceptable personal circumstances must be requested by email in advance of the deadline. 6. For late submissions (i.e. without an agreed extension), a penalty of 5% will be applied for every day a submission is late. 7. No submissions will be accepted after Friday November 8th 2019 @ 23:59 unless an extension has been agreed.   **NB**: Anything submitted later than this date without agreement will be ignored.   1. Assignments which do not adhere to the requirements or which are submitted incorrectly will attract a penalty of up to 10%. 2. No resubmission of assignments after feedback is given is allowed. |

|  |
| --- |
| DESCRIPTION You are expected to:   * State the concepts you are interested in; * Present a summary of the data used, critically discussing relevant issues which impact statistical analysis; * State five (5) hypotheses that you can test to investigate correlation and difference including:   + At least one involving correlation;   + At least one involving difference involving a categorical variable with 2 values;   + At least one involving a categorical variable with more than 2 values; * Conduct appropriate statistical tests to test your hypotheses using R; * Present the findings of the statistical tests used adopting APA guidelines; * Interpret the findings for the stated hypotheses; * You should cite appropriate sources (which are accessible) in order to support your decision making and interpretation of findings and report these using APA guidelines.   You will need to demonstrate:   * An ability to generate and correctly state hypotheses; * The ability to correctly analyse, present and critically assess the dataset used from the perspective of statistical analysis; * The ability to correctly execute, present and interpret appropriate statistical tests for correlation and difference using statistical software; * The ability to interpret the findings gained from your statistical analysis in a clear and accurate way; * The ability to report on the outcomes of statistical tests; * The ability to interpret outcomes of statistical tests and report on this interpretation in the context of a statistical inquiry. |

|  |
| --- |
| DELIVERABLES  * You are required to address two aspects in your submission:   + A report constructed adhering to APA guidelines which addresses the following:     - State clearly the hypotheses you intend to test.       * At least five hypotheses are required:         + At least one involving correlation;         + At least two involving difference one of which must involve a categorical variable with more than 2 values;     - Analyse and describe your variables of interest:       * You must describe your variables in terms of their statistical measurement types and describe them with appropriate descriptive statistics and graphs.       * You must address all issues which could impact on the choice of statistical tests.     - Justify your choice of statistical test based on your findings.     - Report the outcomes of the tests conducted in paragraphs using full sentences using APA style for reporting statistical results.       * Justify your choice of test based on your assessment of the dataset.       * Comment on effect as well as statistical significance.     - Interpret your findings appropriately relevant to your hypotheses. Report on them correctly   + The format of your report is at your discretion.   + A useful guide to creating a report of a statistical inquiry using APA guidelines is available at <http://www.discoveringstatistics.com/docs/writinglabreports.pdf>.   + The R commands plus output generated from this to support the statistics and information included in your report is required. It should be possible to execute the R commands to verify the statistics you have included. |

|  |
| --- |
| SUBMISSION All required documents should be submitted using the **CA Part I Assignment** in Brightspace.   * You must include the following information at the start of all files submitted:   + Student Number: <<your student number>>   + Student Name: <<your name>>   + Programme Code: <<programme code>>   + The version of R used.   + The R packages needed for your code to execute successfully. * All files must include your student number at the start of the file name e.g. D123456.rmd, D123455.nb.html.   You have choices for your submission:   * Option A: R notebook which includes the commands and creates html with the nb.html created from this. * Option B: A pdf file including all required reporting plus an R script well commented to indicate which sections of the report commands relate to plus an output file (html, pdf, word) that includes the output from these statistical tests well commented so that the commands that generated the commands can be found. |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Correct statement of hypotheses. | 15 | | Inspection, presentation and assessment and presentation of variables of interest. | 20 | | Correct identification conduct and reporting on correlation and difference tests. | 50 | | Valid interpretation of the findings for hypotheses and question. | 15 | |  | 100 |  BASIC MARKING SCHEME |