# Higher Diploma in Data Analytics (HDSDA)
# Programming for Big Data Project Description
# Semester 1
# 2019 - 2020

You are required to carry out a series of analyses on publicly accessible datasets using the R programming language used in this module and programming environments suitable for the task. It is recommended that your use at least two separate datasets. For each of the chosen datasets you are required to compile a report of your analysis. Each dataset should have at least 1,000 records (rows). If you are unsure if your dataset(s) is/are appropriate, please check with your lecturer. You must provide evidence in your report that you are authorized to use the dataset(s) that you have chosen.

The main deliverable is a report that provides significant insights into the datasets that you have chosen to analyse. Your report should provide at least **four unique insights** based on your data analysis. Examples of insights might include relationships, trends/patterns, correlations, models based on the data, visuals, and statistical analyses.

All deliverables should be compiled into a project report document for submission along with all programming code elements in an appendix. Please submit your report via the Turnitin upload link in Moodle. R scripts and additional files are to be uploaded to a separate link in Moodle. Your project report should discuss the challenges that you encountered while handling your chosen datasets and the means and mechanisms you implemented to overcome these challenges. The word count for your report should be not less than 2,000 words, and not more than 2,500 words (not counting R code).

## Structure and Rating Grid

- Description of the objective(s) of the analysis with reference to basic domain literature to explain the domain purpose of the analyses **[10%]**

- Description of the underlying dataset including an assessment of the data types present, with an emphasis on the data that is actually used in the analytical processes **[15%]**

- Approach to the analysis, aided by visuals such as diagrams, flowcharts and tables where appropriate **[5%]**

- R code demonstrating at least four unique insights. R scripts will be executed as part of the assessment process. It is expected that scripts are fully working, efficient, commented clearly, and do not contain excess code **[50%]**

- Project report structure, presentation and discussion of challenges. **[20%]**

**Note:** the project contributes towards a maximum of 50% of the marks for the module.

# HDSDA Assessment Rubric – Programming for Big Data Project

| | Solid H1 (≥ 80%) | H1 (≥ 70%) | H2.1 (≥ 60%) | H2.2 (≥ 50%) | Pass (≥ 40%) | Fail (< 40%) |
|---|---|---|---|---|---|---|
| Description of the objective(s) of the analysis with reference to basic domain literature to explain the domain purpose of the analyses.<br><br>(10%) | Very challenging project objectives are well presented, met, and thoroughly discussed.<br><br>Excellent critical analysis of substantive and relevant literature. | Challenging project objectives are well presented, met, and thoroughly discussed.<br><br>Very good analysis of substantive and relevant literature. | Reasonable project objectives are well presented, met, and discussed.<br><br>Good analysis of relevant literature. | Reasonable project objectives are clear, and are mostly met.<br><br>Adequate analysis of mostly relevant literature. | There are clear objectives, which are at least partially met.<br><br>Some review of some relevant literature. | Cannot discern project objectives, and/or if project objectives were met.<br><br>No relevant literature reviewed. |
| Description of the underlying dataset including an assessment of the complexity and data types present.<br><br>(10%) | The datasets have been very well described, prepared, and meaningfully explored.<br>At least two datasets have a high degree of complexity.<br>The project significantly exceeds the stated minimum requirements. | The datasets have been well prepared and meaningfully explored.<br><br>At least two datasets have a high degree of complexity. | The datasets have been well prepared and explored.<br><br>At least one dataset has a high degree of complexity. | The datasets have been appropriately prepared for analysis.<br><br>At least one of the datasets is non-trivial. | At least two datasets appropriately handled fitting for the objectives.<br><br>The datasets are probably somewhat trivial. | Less than two datasets.<br><br>No obvious development conducted. |
| Approach to the analysis.<br><br>(10%) | Excellent approach to the analysis, aided by a wide variety of visuals such as pseudocode, diagrams, flowcharts and tables where appropriate. | Very good approach to the analysis, aided by a wide variety of visuals such as pseudocode, diagrams, flowcharts and tables where appropriate. | Good approach to the analysis, aided by a variety of visuals such as pseudocode, diagrams, flowcharts and tables where appropriate. | Good approach to the analysis, aided by some visuals such as pseudocode, diagrams, flowcharts and tables where appropriate. | Basic approach to the analysis, aided by a small number of visuals such as pseudocode, diagrams, flowcharts and tables where appropriate. | Minimal or no appropriate approach to the analysis that is not supported by visuals such as pseudocode, diagrams, flowcharts and tables. |
| Insights and R Code<br><br>(50%)<br><br>(Insights 4 x10%)<br>((Format/Style 10%) | Four or more insights that are presented and thoroughly discussed with appropriate references to existing work.<br><br>R code shows excellent programming techniques such as extensive use of iterative statements, functions, and modelling.<br><br>Code is fully commented. There are no syntax or logic errors, and no excess code used.<br><br>The implementation significantly exceeds the stated minimum requirements. | Four or more insights that are presented and thoroughly discussed with appropriate references to existing work.<br><br>R code shows excellent programming techniques such as extensive use of iterative statements, functions, and modelling.<br><br>Code is fully commented. There are no syntax or logic errors, and no excess code used. | Four or more insights that are presented and discussed with some references to existing work.<br><br>R code shows very good programming techniques such as extensive use of iterative statements, functions, and modelling.<br><br>Code is partially commented. There are few syntax or logic errors, and a minimal amount of excess code used. | Four or more insights that are presented and partially discussed with few references to existing work.<br><br>R code shows good programming techniques such as partial use of iterative statements, functions, and modelling.<br><br>Code is partially commented. There are several syntax or logic errors, and use of excess code. | Four or more insights with basic presentation and discussion with little reference to existing work.<br><br>R code shows poor programming techniques and little use of iterative statements, functions, and modelling.<br><br>Code is poorly commented. There are many syntax or logic errors, and excess use of unnecessary code. | Less than four insights with poor presentation and discussion with no reference to existing work.<br><br>R code shows very poor programming techniques with minimal use of iterative statements, functions, and modelling.<br><br>Code is barely commented. Very few lines of code work and there are many syntax or logic errors, and excess use of much unnecessary code. |
| Project report structure, presentation and discussion of challenges (20%) | Well written, with no (large) language errors. All figures are well conceived and readable.<br><br>There is significant reflection on the challenges faced in this project.<br><br>Paper presented in scientific format (eg, IEEE).<br><br>The report significantly exceeds the stated minimum requirements. | Well written, with no (large) language errors. All figures are well conceived and readable.<br><br>There is significant reflection on the challenges faced in this project.<br><br>Paper presented in scientific format (eg, IEEE). | Main document has a few language and/or style errors.<br><br>There is very good reflection on the challenges faced in this project.<br><br>Paper is presented in a well-structured format.<br><br>Visuals are clear and easy to read. | Main document has a some language and/or style errors.<br><br>There is good reflection on the challenges faced in this project.<br><br>Paper is presented in a structured format.<br><br>Visuals are somewhat clear and easy to read. | Main document has a several language and/or style errors.<br><br>There is some reflection on the challenges faced in this project.<br><br>Paper is presented in a poorly structured format.<br><br>Visuals are not clear or easy to read. | Main document has a multiple language and/or style errors.<br><br>There is no reflection on the challenges faced in this project.<br><br>Paper is presented in a very poor or no structured format.<br><br>Visuals are poor and difficult to read. |