CCT College Dublin Continuous Assessment

Programme Title:	BSc (Hons) in Computing in IT (4th Yr)		
Cohort:	FT		
Module Title(s):	Data Exploration & Preparation		
Assignment Type:	Pair (Max 2 students)	**Weighting(s)**:	40% (60% pair group and 40% individual)
Assignment Title:	CA1 Project		
Lecturer(s):	Dr. Muhammad Iqbal		
Issue Date:	9th October 2023		
Submission Deadline Date:	3rd December 2023		
Late Submission Penalty:	Late submissions will be accepted up to **5** calendar days after the deadline. All late submissions are subject to a penalty of **10%** of the mark awarded. Submissions received more than 5 calendar days after the deadline above **will not** be accepted and a mark of 0% will be awarded.		
Method of Submission:	**Moodle**		
Instructions for Submission:	Upload all files MS word file, jupyter notebook, dataset and any supporting information on Moodle.		
Feedback Method:	**Results posted in Moodle gradebook**		
Feedback Date:	3 weeks after submission		

Learning Outcomes:
Please note this is not the assessment task. The task to be completed is detailed on the next page.
This CA will assess student attainment of the following minimum intended learning outcomes:

1. Develop strategies for identifying and handling missing and out-of-range data, as well as feature engineering as part of the preparation phase of data analysis. (Linked to PLO 4 (Stage 4 SLO 4))
2. Understand the purpose of and methods to achieve dimensionality reduction and the difference between dimensionality reduction and feature selection. (Linked to PLO 1 / PLO 3 (Stage 4 SLO 1 / SLO 3))
3. Select and perform appropriate feature selection and/or dimensionality reduction techniques on a variety of wide datasets. (Linked to PLO 3 (Stage 4 SLO 3))

Attainment of the learning outcomes is the minimum requirement to achieve a Pass mark (40%). Higher marks are awarded where there is evidence of achievement beyond this, in accordance with QQI *Assessment and Standards, Revised 2013*, and summarised in the following table:

Percentage Range	**CCT Performance Description**	**QQI Description of Attainment**	
		Level 6, 7 & 8 awards	**Level 9 awards**

90% +	Exceptional	Achievement includes that required for a Pass and in **most** respects is significantly and consistently beyond this	Achievement includes that required for a Pass and in **most** respects is significantly and consistently beyond this
80 – 89%	Outstanding		
70 – 79%	Excellent		
60 – 69%	Very Good	Achievement includes that required for a Pass and in **many** respects is significantly beyond this	Achievement includes that required for a Pass and in **many** respects is significantly beyond this
50 – 59%	Good	Achievement includes that required for a Pass and in **some** respects is significantly beyond this	Attains all the minimum intended programme learning outcomes
40 – 49%	Acceptable	Attains all the minimum intended programme learning outcomes	
35 – 39%	Fail	Nearly (but not quite) attains the relevant minimum intended learning outcomes	Nearly (but not quite) attains the relevant minimum intended learning outcomes
0 – 34%	Fail	Does not attain some or all of the minimum intended learning outcomes	Does not attain some or all of the minimum intended learning outcomes

Please review the CCT Grade Descriptor available on the module Moodle page for a detailed description of the standard of work required for each grade band.

The grading system in CCT is the QQI percentage grading system and is in common use in higher education institutions in Ireland. The pass mark and thresholds for different grade bands may be different from what you have experience of in the higher education system in other countries. CCT grades must be considered in the context of the grading system in Irish higher education and not assumed to represent the same standard the percentage grade reflects when awarded in an international context.

Assessment Task

This is a pair-based project (Max 2 students) using R programming language or any other language of your choice. Analyse a specific problem only in the following areas,

- Crime
- Covid 19
- Dublin Transport

The dataset should have at least 7000 rows and 10 columns after cleaning and there is not any upper bound. The type of question(s) that you should formulate for the project will depend on the chosen domain of the dataset that your pair is considering for the Data Exploration and Preparation (DEP) project. The objectives of the DEP project are based on the domain knowledge of data. The pair would need to complete the following tasks during the development of this pair project.

a) Identify which variables are categorical, discrete and continuous in the chosen data set and show using some visualization or plot. Explore whether there are missing values for any of the variables.
b) Calculate the statistical parameters (mean, median, minimum, maximum, and standard deviation) for each of the numerical variables.
c) Apply Min-Max Normalization, Z-score Standardization and Robust scalar on the numerical data variables.
d) Line, Scatter and Heatmaps can be used to show the correlation between the features of the dataset.
e) Graphics and descriptive understanding should be provided along with Data Exploratory analysis (EDA). Identify subgroups of features that can explore some interesting facts.
f) Apply dummy encoding to categorical variables (at least one variable used from the data set) and discuss the benefits of dummy encoding to understand the categorical data.
g) Apply PCA with your chosen number of components. Write up a short profile of the first few components extracted based on your understanding.
h) What is the purpose of dimensionality reduction? Explore the situations where you can gain the benefit of dimensionality reduction for data analysis.

Your pair will present their findings and defend the results in the report (MS Doc/ pdf or any other readable format). Your report should capture the following aspects that are relevant to your project investigations.

i) Description of problem domain, motivation, data set chosen and challenges faced during this project. Your pair should provide the characterization, description and explanation of techniques used to prepare the data set (size / attributes / missing values / outliers).

(15 marks)

ii) Find unusual patterns by identifying variations and covariation between the features in the dataset and perform Exploratory Data Analysis (EDA) to justify outcomes with supporting questions and visualizations.

(20 marks)

iii) Show the implementation of an encoding scheme, such as one-hot, Label etc. Apply Principal Component Analysis (PCA) for the dimensionality reduction on the chosen dataset. Interpret and explain the outcomes obtained using PCA.

(15 marks)

iv) Provide an explanation of the code submitted along with the project (Code must be commented). Conclusions of the project should be specified at the end of the report. Citations and references to be in Harvard Style.

(10 marks)

v) Each team member presents a PowerPoint presentation of their work (maximum 5 slides) to emphasize their distinctive contributions based on their involvement in the project's conceptual understanding, code development, and deployment.

(20 marks individual)

vi) Each team member fully described their individual contributions to the project in a reflective journal, using at least 400 to 500 words as well as images, diagrams, figures, and visualizations to elaborate his/ her work.

(20 marks individual)

Submission Requirements

All assessment submissions must meet the minimum requirements listed below. Failure to do so may have implications for the marks awarded.

- The code and datasets should be provided and uploaded on Moodle.
- Must be clearly specified the number of words used in the report.
- Number of Words in the report for pair report (Min: 2500 words and Max: 3000) excluding diagrams and code.
- If you are doing it as an individual, the total number of words is 1500 words for this CA1.
- Describe the contribution of each team member in the project clearly and use a bar chart or pie chart to represent the effort and time spent during this project.
- The rubric is provided for the detailed breakdown of marks at the end of this CA.
- Use Harvard Referencing when citing third party material
- Be the student's own work.
- Include the CCT assessment cover page.
- Be submitted by the deadline date specified or be subject to late submission penalties
- Note: The names of pair members must be uploaded on the link provided on Moodle until 15th October 2023 (23:59).
- Describe the contribution of each team member in the project clearly and use a bar chart or pie chart to represent the effort and time spent during this project. Use version control like Github or any other tool to show the progress of both team members in CA1. You should have at least 5 commits on Github before submission.

GRADING RUBRIC – Data Exploration and Preparation – 2023 - 24								
GRADE	**90-100%**	**80-90%**	**70-79%**	**60-69%**	**50-59%**	**40-49%**	**35-39%**	**<35%**
Performance	**Exceptional**	**Outstanding**	**Excellent**	**Very Good**	**Good**	**Acceptable**	**Fail**	**Fail**
Introduction to problem Description, Motivation and Characterization and Description (15%)	An exceptional introduction to problem description and motivation, characterization and cleaning of a dataset that provide a concise and clear case for the proposed Data Exploration and Preparation project.	An outstanding introduction to problem description and motivation, characterization and cleaning of a dataset that provide a compact and clear case for the proposed Data Exploration and Preparation project.	An excellent introduction to problem description and motivation, characterization and cleaning of a dataset that provide a precise and clear case for the proposed Data Exploration and Preparation project.	A very good introduction to problem description and motivation, characterization and cleaning of a dataset that provides a very convincing case for the proposed Data Exploration and Preparation project.	A good introduction to problem description and motivation, characterization and cleaning of a dataset that furnishes a largely convincing case for the proposed Data Exploration and Preparation Project.	An adequate introduction to problem description and motivation, characterization and cleaning of a dataset that offers a somewhat weak case for the proposed Data Exploration and Preparation Project.	A poor introduction to problem description and motivation that fails to motivate, clean and characterise the dataset for the problem or provide a case for the proposed Data Exploration and Preparation Project.	An impecunious introduction to problem description that fails entirely to motivate the problem.
EDA and unusual patterns (20%)	An exceptional strategy is implemented to perform EDA by identifying variations and covariation between the features in the dataset to justify outcomes. Use of appropriate visualizations.	An outstanding strategy is employed to perform EDA by identifying variations and covariation between the features in the dataset to justify outcomes. Use of nice visualizations.	An excellent strategy is considered to perform EDA by identifying variations and covariation between the features in the dataset to justify outcomes. Use of proper visualizations.	A very good strategy is used to perform EDA by identifying variations and covariation between the features in the dataset to justify outcomes. Use of very good visualizations.	A good strategy is applied to perform EDA by identifying variations between the features. Use of good visualizations.	An adequate strategy is partially used to perform EDA. Use of visualizations.	A poor strategy is used to perform EDA. No visualizations.	An impecunious strategy is provided and No visualizations.
Interpretation of results using PCA (15%)	An exceptional interpretation and explanation of the results based on problem specification and objectives. The results clearly exhibit the use of PCA and encoding schemes. An exceptional justification is provided.	An outstanding interpretation and explanation of the results based on problem specification and objectives. The results clearly exhibit the use of PCA and encoding schemes. An outstanding advocacy is provided.	An excellent interpretation and explanation of the results based on problem specification and objectives. The results clearly exhibit the use of PCA and encoding schemes. An excellent defence is provided.	A very good interpretation and explanation of the results based on problem specification and objectives. The results clearly exhibit the use of PCA and encoding schemes. A very good justification is provided.	A good interpretation and explanation of the results based on problem specification and objectives. The results exhibit the use of PCA and encoding schemes. A good justification is provided.	An adequate interpretation of the results based on problem specification and objectives. The results exhibit the partial use of PCA and encoding schemes. An adequate justification is provided.	A poor interpretation of the results based on problem specification and objectives. No clear use of PCA and encoding schemes.	An impecunious interpretation of the results. No use of PCA and encoding schemes.
Code description and comments, Conclusions, citations, and references (10%)	An exceptional description of code using logical comments. The comments are detailed and provide a remarkable understanding of the functionality of the code. An exceptional report along with proper	An outstanding description of code using rational comments. The comments are detailed and provide an impeccable understanding of the functionality of the code. An outstanding	An excellent description of code using comments. The comments are detailed and provide an explicit understanding of the functionality of the code. An excellent report along with proper conclusion, citations and	A very good description of code using comments. The comments are brief and provide a clear understanding of the functionality of the code. A very good report along with proper conclusion, citations and	A good description of code using comments. The comments are very brief and provide an understanding of the functionality of the code. A good report along with proper conclusion, citations and	An adequate description of code using comments. The comments are not satisfactory and provide a partial understanding of the functionality of the code. An adequate report along with proper conclusion, citations and	A poor description of code using comments. The comments are not satisfactory. A poor report along with proper conclusion, citations and references in all sections.	An impecunious code using unsatisfactory comments. The report is not in acceptable format and poorly designed and written.

	conclusion, citations and references in all sections.	report along with proper conclusion, citations and references in all sections.	references in all sections.	references in all sections.	references in all sections.	references in all sections.		
Powerpoint presentation (20%) - Individual	The presentation is delivered in an exceptional manner, is well-organized and visually appealing, and successfully explains the topic's essential concepts, ideas, and code.	The presentation is delivered in an outstanding manner, is well-organized and visually appealing, and successfully explains the topic's essential concepts, ideas and code.	The presentation is delivered in an excellent manner, is well-organized and visually appealing, and successfully explains the topic's essential concepts, ideas, and code.	The presentation is delivered in a very good manner, is nicely organized and visually appealing, and decently explains the topic's essential concepts, ideas and code.	The presentation is delivered in a good manner, is organized and visually appealing, and explains the topic's essential concepts, ideas, and code.	The presentation is delivered in an acceptable manner, is organized, and explains the topic's essential concepts, ideas, and code to some extent.	The presentation is delivered in a poor manner, is not organized, and has an unsuccessful explanation of the topic's concepts, ideas, and code.	The presentation is not delivered according to the guidelines.
Reflective journal for individual pair member (20%) - Individual	Reflection demonstrates an exceptional level of engagement and understanding of the pair project material, and shows exceptional evidence of critical thinking, self-reflection, and collaboration.	Reflection demonstrates an outstanding level of engagement and understanding of the pair project material, and shows outstanding evidence of critical thinking, self-reflection, and collaboration.	Reflection demonstrates an excellent level of engagement and understanding of the pair project material, and shows excellent evidence of critical thinking, self-reflection, and collaboration.	Reflection demonstrates a very good level of engagement and understanding of the pair project material, and shows very good evidence of critical thinking, self-reflection, and collaboration.	Reflection demonstrates a good level of engagement and understanding of the pair project material, and shows good evidence of critical thinking, self-reflection, and collaboration.	Reflection demonstrates an acceptable level of engagement and understanding of the pair project material, and shows some evidence of critical thinking, self-reflection, and collaboration.	Reflection demonstrates a poor level of engagement and understanding of the pair project material, and shows incomplete evidence of critical thinking, self-reflection, and collaboration.	Reflection does not demonstrate any engagement and understanding of the pair project material, and shows no evidence of critical thinking, self-reflection, and collaboration.