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:	Group	**Weighting(s)**:	40%
Assignment Title:			
Lecturer(s):	Dr. Muhammad Iqbal		
Issue Date:	14th October 2022		
Submission Deadline Date:	11th December 2022		
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 one zip file composed of pdf/ word file, jupyter notebook, dataset and any supporting information.		
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 group-based project (3 - 4 students) using R programming language or any other tool. Analyse a specific problem only in the following areas, such as Hospitals and Patients, Housing, Agriculture, Economy and Environment (The dataset should have at least 10000 rows and 12 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 group is considering for Data Exploration and Preparation (DEP) project. The objectives of DEP project are based on the domain knowledge of data. The group would need to complete the following tasks during the development of this group 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 sub-groups of features that can explore some interesting facts.
f) Apply dummy encoding to categorical variables (at least one variable use 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 group 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 and chosen data set.

(10 marks)

ii) Motivates for the problem and Challenges faced in Data Exploration project.

(10 marks)

iii) Characterization, description and explanation of techniques used to prepare the data set (size / attributes / missing values / outliers).

(10 marks)

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

(30 marks)

v) Show the implementation of encoding scheme, such as one-hot. Apply Principal Component Analysis (PCA) for the dimensionality reduction. Interpret and explain the outcomes obtained using PCA.

(20 marks)

vi) Provide an explanation of the code submitted along with the project (Code must be commented).

(10 marks)

vii) Conclusions of the project should be specified at the end of the report. Citations and references to be in Harvard Style.

(10 marks)

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 in zip format on Moodle.
- Must be clearly specified the number of words used in the report.
- Number of Words in the report (Min: 2000 words and Max: 3500) excluding diagrams and code.
- 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 group members must be uploaded on the link provided on Moodle until 25th October 2022 (23:59).
- Must be clearly specified the number of words used after each section in the report.

GRADING RUBRIC – Data Exploration and Preparation – 2022 - 23								
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 and Motivation (10%)	An exceptional introduction to problem description and motivation that provide a concise and clear case for the proposed Data Exploration and Preparation project.	An outstanding introduction to problem description and motivation that provide a compact and clear case for the proposed Data Exploration and Preparation project.	An excellent introduction to problem description and motivation that provide a precise and clear case for the proposed Data Exploration and Preparation project.	A very good introduction to problem description and motivation that provide offers a very convincing case for the proposed Data Exploration and Preparation project.	A good introduction to problem description and motivation that furnishes a largely convincing case for the proposed Data Exploration and Preparation Project.	An adequate introduction to problem description and motivation 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 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.
Project Objectives (10%)	An exceptional specification of objectives concisely.	An outstanding specification of objectives precisely.	An excellent specification of objectives succinctly.	A very good specification of objectives.	A good specification of objectives.	An adequate specification of objectives.	A poor specification of objectives.	An impecunious specification of objectives.
Characterization and Description along with cleaning of Dataset (10%)	An exceptional characterization and cleaning of dataset that abstracts all details from source to fields.	An outstanding characterization and cleaning of dataset that highlights all details from source to fields.	An excellent characterization and cleaning of dataset that summarizes all details from source to fields.	A very good characterization and cleaning of dataset that summarizes all details from source to fields.	A good characterization and cleaning of dataset that summarizes all details from source to fields.	An adequate characterization and cleaning of dataset that summarizes all details from source to fields.	A poor characterization and cleaning of dataset that summarizes all details from source to fields.	An impecunious characterization and cleaning of dataset.
EDA and unusual patterns (30%)	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 (20%)	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 (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 outstanding description of code using rational comments. The comments are detailed and provide an impeccable understanding of the functionality of the code.	An excellent description of code using comments. The comments are detailed and provide an explicit understanding of the functionality of the code.	A very good description of code using comments. The comments are brief and provide a clear understanding of the functionality of the code.	A good description of code using comments. The comments are very brief and provide an understanding of the functionality of the code.	An adequate description of code using comments. The comments are not satisfactory and provide a partial understanding of the functionality of the code.	A poor description of code using comments. The comments are not satisfactory.	An impecunious code using unsatisfactory comments.
Conclusions, citations, and references (10%)	An exceptional presentation of conclusions. An exceptional report along with proper citations and references in all sections.	An outstanding manifestation of conclusions. An outstanding report along with appropriate citations and references in all sections.	An excellent demonstration of conclusions. An excellent report along with proper citations and references in all sections.	A very good demonstration of conclusions. A very good report along with proper citations and references in all sections.	A good demonstration of conclusions. A good report along with citations and references in some sections.	An adequate demonstration of conclusions. An adequate report along with incomplete citations and references.	A poor demonstration of conclusions or no conclusions. A report along with errors.	An impecunious demonstration of conclusions or no conclusions. An inadequate report.