# MSC DA CA1

|  |  |
| --- | --- |
| Created By: | Stephen Kelly |
| Created On: | 24/10/2023 |
| Updated By: | Stephen Kelly |
| Last Update: | 24/10/2023 |

Contents

[MSC DA CA1 1](#_Toc149043729)

[Assessment Task 3](#_Toc149043730)

[Criteria of Analysis 4](#_Toc149043731)

[Statistics for Data Analytics - Criteria of Analysis 4](#_Toc149043732)

[Statistics: (Graded out of 100 & 1200 words) 4](#_Toc149043733)

[Data Preparation & Visualisation - Criteria of Analysis 5](#_Toc149043734)

[Data preparation and Visualization : (Graded out of 100 & 1200 words) 5](#_Toc149043735)

[Machine Learning for Data Analysis - Criteria of Analysis 6](#_Toc149043736)

[Machine learning for Data Analytics:(Graded out of 100 & 1200 words) 6](#_Toc149043737)

[Programming for DA - Criteria of Analysis 7](#_Toc149043738)

[Programming**:** (Graded out of 100 & 600 words) 7](#_Toc149043739)

# Assessment Task

Students are advised to review and adhere to the submission requirements documented after the assessment task.

**Scenario:** Population in Ireland

A large amount of data has been collected by The Central Statistics Office in Ireland in relation to the population of Ireland, This data is available at:

<https://data.cso.ie/product/pme>

Chosen File: Population Estimates (Persons in April)

<https://data.cso.ie/>

You are required to choose a particular area of interest and formulate the appropriate questions for modelling and analysis. For Example (but not limited to):

* Annual Population Change
* Immigration and Migration
* Population Forecasting
* etc…

You are required to collect, process, analyse and interpret the data in order to identify possible issues/problems at present and make predictions/classifications in regard to the future. This analysis will rely on the available data from CSO and **any additional data you deem necessary** (with supporting evidence) to support your hypothesis for this scenario.

This will require you to employ critical analysis of not only the domain of choice but also for the regression and or classification that you undertake.

**Note: This is an academic exercise and not a hypothetical report to the CSO.**

# Criteria of Analysis

**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:

## Statistics for Data Analytics - Criteria of Analysis

1. Explore and evaluate datasets using descriptive statistical analyses. (PLO 1)
2. Apply statistical analysis to appropriate datasets and critique the limitations of these models (PLO 2,4)
3. Utilise current software tools and languages to produce and document result sets from existing data (e.g., spreadsheets, R, Python). (PLO 1,4)

## Statistics: (Graded out of 100 & 1200 words)

You need to analyse the chosen dataset using statistical logic and statistical techniques. Note: ALL Statistical work MUST be carried out using Python.

You are required to:

1. Summarise your dataset clearly, using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You should critically analyse your findings, in addition to including the necessary Python code, output and plots in the report. You are required to plot at least three graphs. [0-35]
2. Use two discrete distributions (Binomial and/or Poisson) in order to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualise your data and explain what happens with the large samples in these cases. You must work with Python and your mathematical reasoning must be documented in your report. [0-30]
3. Use Normal distribution to explain or identify some information about your dataset. [0-20]
4. Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case. [0-15]

**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:

## Data Preparation & Visualisation - Criteria of Analysis

1. Discuss the concepts, techniques and processes underlying data visualisation to critically evaluate visualisation approaches with respect to their suitability for different problem areas. (linked to PLO 1)
2. Programmatically Implement graphical methods to identify issues within a data set (missing, out of range, dirty data) (linked to PLO 3, PLO 5)
3. Engineer new features selection in data with the goal of improving the performance of machine learning models. (linked to PLO 2, PLO 4)

## Data preparation and Visualization : (Graded out of 100 & 1200 words)

1. You must perform appropriate EDA on your dataset, rationalizing and detailing why you chose the specific methods and what insight you gained. [0-20]
2. You must also rationalise justify and detail all the methods used to prepare the data for ML. [0-30]
3. Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. [0-20]
4. All design and implementation of your visualizations must be justified and detailed in full., making reference to Tufts Principles [0-30]

**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:

## Machine Learning for Data Analysis - Criteria of Analysis

1. Develop a machine learning strategy for a given domain and communicate effectively to team members, peers and project stakeholders the insight to be gained from the interpreted results. (Linked to PLO 1, PLO 4, PLO 6)
2. Implement a range of classification and regression techniques and detail / document their suitability for a variety of problem domains. (Linked to PLO 5)
3. Critically evaluate the performance of Machine Learning models, propose strategies to optimise performance. (Linked to PLO 3)

## Machine learning for Data Analytics:(Graded out of 100 & 1200 words)

1. Explain which project management framework (CRISP-DM, KDD or SEMMA) is required for a data science project. Discuss and justify with real-life scenarios. Provide an explanation of why you chose a supervised, unsupervised, or semi-supervised machine learning technique for the dataset you used for ML modelling. **[0 - 20]**
2. Machine learning models have a wide range of uses, including prediction, classification, and clustering. It is advised that you assess several approaches (at least two), choose appropriate hyperparameters for the optimal outcomes of Machine Learning models using an approach of hyperparameter tunning, such as GridSearchCV or RandomizedSearchCV. **[0 - 30]**
3. Show the results of two or more ML modelling comparisons in a table or graph format. Review and critically examine the machine learning models' performance based on the selected metric for supervised, unsupervised, and semi-supervised approaches. **[0 - 30]**
4. Demonstrate the similarities and differences between your Machine Learning modelling results using the tables or visualizations. Provide a report along with an explanation and interpretation of the relevance and effectiveness of your findings. **[0 - 20]**

**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:

## Programming for DA - Criteria of Analysis

1. Debate the selection of programming concepts in the design of programmatic solutions, in terms of paradigm and language selection. (PLO 1).
2. Design and implement algorithms for use within the context of DA. (PLO 2).

## Programming**:** (Graded out of 100 & 600 words)

1. The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

1. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**