A picture containing icon

Description automatically generated

**CCT College Dublin Continuous Assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Programme Title:** | MSc in Data Analytics | | |
| **Delivery Mode:** | FT/SB+ | | |
| **Cohort Details:** | *MSc in Data Analytics (Sept24 start) Stage 1 Sem 1* | | |
| **Module Title(s)**: | *Programming for DA*  *Stats for Data Analytics*  *ML for Data Analysis*  *Data Prep & Vis*  *Programme schedules are all published on the* [*CCT IQR Provider Profile*](https://irq.ie/providers/cct-college-dublin?id=fec9ea7a-ace4-42c7-9fd5-7fccb6f0a53a&ref=%257B%2522search%2522:%2522cct%2522%257D) | | |
| **Assignment Type:** | *Individual* | **Weighting(s):** | *Programming for DA 50%*  *Stats for Data Analytics 50%*  *ML for Data Analysis 50%*  *Data Prep & Vis 50%* |
| **Assignment Title:** | *MSC\_DA\_CA1* | | |
| **Lecturer(s)**: | *Taufique Ahmed/Cristiani Eccher*  *Kayoum Khbuli*  *Muhammad Iqbal*  *David McQuaid* | | |
| **Issue Date:** | *1st October 2024* | | |
| **Submission Deadline Date:** | *8th November 2024* | | |
| **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**  **Use the submission link on the Data Visualisation and Preparation Module page** | | |
| **Instructions for Submission:** | *Please do not ZIP your files. ALL files must be uploaded individually (to a maximum of 20 files)*  *Expected files : Written report (word document only, NO PDF’s) ,Code files (Jupyter notebook (.ipynb) ONLY, NO PYTHON FILES), Data Files*  *Note that the maximum number of Jupyter Notebooks is 4* | | |
| **Feedback Method:** | **Results posted in Moodle gradebook** | | |
| **Feedback Date:** | *[Insert date provisional result and feedback will be released to students in the appropriate format e.g. Monday, dd-mm-yyyy]* | | |

### Assessment Outline

### Description of Assessment Task

### Scenario: Tourism in Ireland

## A large amount of data has been collected in relation to Tourism in Ireland, Examples of this data is available at:

## <https://www.cso.ie/en/statistics/tourismandtravel/>

## <https://www.failteireland.ie/Research-and-Insights.aspx>

## <https://www.tourismireland.com/research-and-insights#/>

**Note you may also use any other RELEVENT datasets, but you may NOT use data from KAGGLE.**

## 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 Tourism Change

## Local Tourism

## Tourist Accommodation

## 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 (see above) 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 statistical, data preparation/visualization, programming, and regression and/or classification that you undertake.

## Note: This is an academic exercise and not a hypothetical report and the most important aspect of this report is evaluating and rationalizing your decisions in the domain of Data Analytics NOT the problem domain.

## All Project files MUST be uploaded into MOODLE, this is your responsibility, if any files are not uploaded to MOODLE, even if they are available on GITHUB, they will NOT BE GRADED.

## Criteria of Analysis (ALL EXPERIMENTAL WORK MUST BE CARRIED OUT USING PYTHON IN JUPYTER NOTEBOOK)

## Statistics: (Graded out of 100)

## 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. Summarize 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]

## 3. Use two discrete distributions (Binomial and/or Poisson) to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualize 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]

## 4. Use Normal distribution to explain or identify some information about your dataset. [0-20]

## 5. 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]

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

## 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 (Scaling, Encoding, imputation etc…). [0-40]

## 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., referring to Tufts Principles [0-20]

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

1. Explain which project management framework (CRISP-DM, KDD or SEMMA) is suitable for a data science project. Discuss and justify your selection with real-life scenarios. Provide an explanation of why you select 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 should choose at least two machine learning models, find appropriate hyperparameters to demonstrate the optimal outcomes of Machine Learning models using an approach of hyperparameter tuning, such as GridSearchCV or RandomizedSearchCV or any other approach. [0 - 30]

3. Present a comparison of the results from at least two machine learning models using a clear and well-organized table or graph. Critically evaluate the performance of each model, focusing on the selected metrics relevant to supervised, unsupervised, or semi-supervised approaches. [0 - 30]

4. Logically interpret and explain the similarities and differences between the results obtained from the machine learning models. Provide a detailed report discussing the relevance, effectiveness, and practical implications of your findings, demonstrating a clear understanding of the models' performance and their suitability for the task. [0 - 20]

## Programming : (Graded out of 100)

## YOU MUST ATTEMPT BOTH TASKS

## 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).

## 2. In a dedicated section in your report, 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.

## 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.

## Assessment Requirements

Note ALL Students are required to use Git Classroom for any Assignments that they are working on. This assignments Git Classroom link is <https://classroom.github.com/a/btwmBr_P>

This means that ALL changes must be committed to the assignments Git classroom during your assignment. (Not just a single commit at the end!) This is to allow you to display your incremental progress throughout the assessments, give you practice for your capstone/thesis, allows you to create an online portfolio that can be used to showcase your work and to ensure that there are no problems with final uploads (as all your work will be available on GitHub). It is expected that there will be a minimum of 10 commits (with many of you making very many more).

You may Only use your CCT email for your git account, private/work email-based accounts will not be accepted. You DO NOT NEED TO include your lecturer's CCT email as a collaborator on your submission as they have automatic access.

**NOTE As well as committing to the Git Classroom you must also upload your work to Moodle as usual for grading.**

All assessment submissions must meet the following minimum requirements:

* Be submitted in the format outlined in the assignment summary table.
* 4000 (+/- 10%) words in report (not including code, code comments, titles, references, or citations)
* Report submission MUST be a word document only (No PDF’s!).
* Code in a Jupyter Notebook file (.ipynb) only but may be referenced in the word document.
* Be submitted by the deadline date specified or be subject to late submission penalties.
* Be submitted via Moodle upload
* Use [Harvard Referencing](http://40.115.124.2/sp/subjects/guide.php?subject=harvardref) when citing third party material.
* Be the student’s own work.
* Include the CCT assessment cover page.

## Learning Outcomes:

This assessment addresses the following module learning outcomes for this module:

**Programming for DA**

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

**Statistics for Data Analytics**

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

**Machine Learning for Data Analysis**

* 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)
* Implement a range of classification and regression techniques and detail / document their suitability for a variety of problem domains. (Linked to PLO 5)
* Critically evaluate the performance of Machine Learning models, propose strategies to optimise performance. (Linked to PLO 3)

**Data Preparation & Visualisation**

* 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)

* Programmatically Implement graphical methods to identify issues within a data set (missing, out of

range, dirty data) (linked to PLO 3, PLO 5)

* Engineer new features selection in data with the goal of improving the performance of machine learning models. (linked to PLO 2, PLO 4)

Statement of Acceptable Use of Artificial Intelligence

|  |
| --- |
| **Acceptable and Unacceptable Use of AI**  *This statement is useful when you are allowing the use of AI tools for certain purposes, but not for others.* |
| * The use of generative AI tools (e.g. ChatGPT, Dall-e, etc.) is permitted in this assignment for the following activities:   + Brainstorming and refining your ideas;   + Fine tuning your research questions;   + Finding information on your topic;   + Drafting an outline to organise your thoughts; and   + Checking grammar and style. * The use of generative AI tools is not permitted in this course for the following activities:   + Impersonating you in classroom context   + Generating code for your assignment   + Writing a draft of a writing assignment   + Writing entire sentences, paragraphs or papers to complete class assignments. * You are responsible for the information you submit based on an AI query. Your use of AI tools must be properly documented and cited. * Any assignment that is found to have used generative AI tools in an unauthorised way will be subject to college disciplinary procedures as outlined in the QA Manual. * When in doubt about permitted usage, please ask for clarification. |

## Grading Criteria Statistics for Data Analytics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criteria** | ***Criteria 1*** You must summarize your dataset clearly, using relevant descriptive statistics and appropriate plots. All plots should be rationalized justified, and clearly presented. You should critically analyze your findings. | ***Criteria 2*** You must show the application of two discrete distributions (Binomial and/or Poisson) and show ability to infer about relevant data under investigation. | ***Criteria 3***  You must apply Normal distribution to explain or identify continuous information, skewness, and overall shape of your dataset | ***Criteria 4***  You must compare the application of distributions used in point 3 and 4 in your analysis. Justify the choice of the variables, skewness and shape of the dataset investigated. |
| **Weighting per criteria** | **35 marks** | **30 marks** | **20 marks** | **15 marks** |
| *Excellent (+70%)* | Comprehensive Descriptive Statistics performed with clear rationale for each method chosen. Insights are deep and well-articulated. | Thorough explanation of both distributions explored, well detailed, and linked to insights. | Thorough explanation of normal distribution explored, well detailed, and linked to insights. | Well-articulated the application of distributions (continuous or discrete) and implementation choices for visualizations are well-justified and detailed. |
| *Very Good (60 - 69%)* | Clear Descriptive Statistics conducted with rationale provided. Insights are relevant and provide good depth | Good explanation of both distributions explored, well detailed, and linked to insights. | Good explanation of normal distribution explored, well detailed, and linked to insights. | Good explanation regarding application of distributions (continuous or discrete) and implementation choices for visualizations are justified and detailed. |
| *Good (50 - 59%)* | Solid Descriptive Statistics conducted with rationale provided. Insights are mostly relevant but may lack depth. | Satisfactory explanation of distributions with adequate justification. Some links to insights but may lack detail. | Satisfactory explanation normal distribution with adequate justification. Some links to insights but may lack detail. | Satisfactory  explanation regarding application of distributions (continuous or discrete) and implementation choices. |
| *Acceptable (40 - 49%)* | Basic Descriptive Statistics conducted, with minimal rationale. Insights are superficial or unclear. | Basic explanation of distributions used; justifications are weak or unclear. | Basic explanation of normal distribution used; justifications are weak or unclear. | Basic explanation regarding application of distributions (continuous or discrete). |
| *Fail (> 39%)* | Inadequate Descriptive Statistics performed; little to no rationale or insight provided. | Little to no explanation of data distributions applied no justifications provided. | Little to no explanation normal distribution applied no justifications provided. | Poor or no justification in comparing distributions from 3 and 4. |

## Grading Criteria Programming for Data Analytics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criteria** | **Criteria 1: Justification of Programming Paradigms and Language Selection** | **Criteria 2: Design and Implementation of Algorithms for Data Analytics** | **Criteria 3: Programmatic Exploration and Use of Python Tools (Code Quality, Libraries, and Jupyter Notebook)** | **Criteria 4: Discussion of Programming Paradigms in the Project and Design Decisions** |
| **Weighting per criteria** | 10 marks | 30 marks | 30 marks | 20 marks |
| *Excellent (+70%)* | Provides a clear, well-supported justification for selecting programming paradigms and languages based on real-life examples. Thoroughly debates choices with strong rationale. | Designs and implements well-structured and efficient algorithms. Includes creative problem-solving techniques, optimizing performance for data analytics tasks. | Demonstrates exceptional code quality, using appropriate Python tools and libraries. Code is well-documented and fully annotated in a Jupyter Notebook. Excellent use of libraries such as Pandas, NumPy, Scikit-learn, etc. | Provides an in-depth discussion of programming paradigms, demonstrating how specific paradigms influenced project design decisions. Examples from the project are clearly linked to paradigms. |
| *Very Good (60 - 69%)* | Offers a strong explanation for selecting programming paradigms and languages. Justification is clear, with relevant real-life examples. | Develops efficient algorithms with appropriate design for data analytics tasks. Some minor areas could benefit from further optimization or refinement. | Demonstrates good use of Python tools and code quality. The project is well-annotated, with minor gaps in documentation. Libraries are used effectively for data analytics tasks. | Provides a well-structured discussion on programming paradigms, showing how they influenced the design. Some minor areas could benefit from deeper explanation or examples. |
| *Good (50 - 59%)* | Provides a reasonable justification for programming paradigm and language choices, though the debate is not deeply developed. | Implements functional algorithms suitable for data analytics. The design is sound but lacks complexity or optimization. | Adequate use of Python tools with reasonably good code quality. Code annotations are present but may lack depth. Libraries are used appropriately but with limited variety or explanation. | Explains the use of programming paradigms in project development, but the connection to design decisions is somewhat basic. Examples from the project are included but lack depth. |
| *Acceptable (40 - 49%)* | Mentions the programming paradigm and language selection with minimal justification or examples. The debate is surface-level and lacks clarity. | Implements basic algorithms, but the design lacks optimization or refinement. Algorithms perform the tasks but may not be efficient. | Python tools and libraries are used but lack thorough explanation or proper documentation. Code quality is basic, and annotations are sparse or unclear. | Discusses programming paradigms but with limited or unclear connection to the project. The discussion lacks examples, and the explanation is brief or generic. |
| *Fail (> 39%)* | Fails to provide a meaningful or clear justification for the programming paradigm and language selection. Real-life examples are missing or irrelevant. | Fails to design or implement appropriate algorithms. The approach is incomplete or incorrect. | Python tools and libraries are poorly used or absent. Code quality is poor, and documentation is incomplete or missing. Code is not well-organized or understandable. | Fails to discuss programming paradigms in a meaningful way. No clear connection is made between the paradigms and the project design decisions. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criteria** | **Criteria 1: Project Management Framework (CRISP-DM, KDD, SEMMA) and Choice of ML Approach/Approaches** | **Criteria 2: Evaluation of Machine Learning Techniques and Hyperparameter Tuning** | **Criteria 3: Comparison, Evaluation and Visualization of ML Modelling Results** | **Criteria 4: Interpretation and Explanation of ML Results** |
| **Weighting per Criteria** | **20 marks** | **30 marks** | **30 marks** | **20 marks** |
| **Excellent (+70%)** | Provides a clear, well-supported justification for the project management framework with detailed real-life examples. Thoroughly explains the rationale for selecting supervised, unsupervised, or semi-supervised ML techniques. | Thoroughly evaluates at least two ML techniques (e.g., classification, prediction, clustering) with comprehensive hyperparameter tuning using GridSearchCV or RandomizedSearchCV. Provides strong reasoning for all decisions. | Compares, evaluates and visualizes at least two ML models using well-structured tables/graphs. Analyses model performance with detailed insights and applies appropriate metrics for each ML approach. | Provides a clear and in-depth interpretation of results. Presents a comprehensive report with meaningful explanations of model effectiveness, performance divergences, and their relevance to the dataset. |
| **Very Good (60 - 69%)** | Provides a strong explanation for the selected project management framework with relevant real-life examples. Justifies the choice of ML technique with a clear connection to the dataset. | Evaluates at least two ML techniques with appropriate hyperparameter tuning. The approach is generally sound, with some minor areas that could benefit from more explanation. | Compares, evaluates and visualizes ML models clearly with organized tables/graphs. Reviews model performance effectively, with well-selected metrics, and minor gaps in analysis. | Explains results with good clarity and insight. Presents a well-structured report with relevant findings, though some areas could be expanded for more depth. |
| **Good (50 - 59%)** | Describes the project management framework with general examples. Provides a reasonable justification for the selected ML technique (supervised, unsupervised, or semi-supervised) in relation to the dataset. | Discusses ML techniques and applies hyperparameter tuning with appropriate methods, though further explanation or details would strengthen the evaluation. | Compares ML models with good tables/graphs and provides some analysis of model performance. Applies metrics at a decent level, though the exploration could be expanded. | Summarizes results with basic explanations and provides some relevant insights. The report presents the findings clearly, though more depth and connection to the dataset would improve the analysis. |
| **Acceptable (40 - 49%)** | Mentions the project management framework with limited explanation or examples. Provides a brief rationale for the selected ML technique but with minimal connection to the dataset. | Attempts evaluation of ML techniques with some hyperparameter tuning. The approach is partially applied but would benefit from clearer rationale and further development. | Provides a basic comparison of ML models with minimal tables/graphs. Performance metrics are applied but require further analysis or organization to support conclusions. | Offers a basic interpretation of results with limited explanations. The report presents the findings but would benefit from a clearer connection to the dataset and more detailed interpretation. |
| **Fail (< 39%)** | Provides an incomplete or unclear explanation of the project management framework and ML technique selection. Real-life examples and rationale are absent or irrelevant to the project. | Fails to provide a structured evaluation of ML techniques or hyperparameter tuning. Models are not adequately chosen or explained. | Fails to provide a meaningful comparison of ML models. Tables/graphs are missing, and metrics are not effectively applied. | Fails to interpret or explain the results. The report is disorganized, with no connection to the dataset or meaningful conclusions. |

## Grading Criteria Machine Learning for Data Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criteria** | ***Criteria 1***  *You must perform appropriate EDA on your dataset, rationalizing and detailing why you chose the specific methods and what insight you gained.* | ***Criteria 2***  *You must also rationalise, justify, and detail all the methods used to prepare the data for ML (Scaling, Encoding, imputation etc…).* | ***Criteria 3***  *Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis.* | ***Criteria 4***  *All design and implementation of your visualizations must be justified and detailed in full., referring to Tufts Principles* |
| **Weighting per criteria** | **20 marks** | **40 marks** | **20 marks** | **20 marks** |
| *Excellent (+70%)* | Comprehensive EDA performed with clear rationale for each method chosen. Insights are deep and well-articulated. | Thorough explanation of all methods used for data preparation (scaling, encoding, imputation, etc.). Justifications are clear, detailed, and linked to insights. | Visualizations are highly relevant, clear, and insightful, significantly enhancing understanding of the dataset. | All design and implementation choices for visualizations are well-justified and detailed, with clear references to Tufts Principles. |
| *Very Good (60 - 69%)* | Clear EDA conducted with rationale provided. Insights are relevant and provide good depth of understanding | Very Good explanation of nearly all methods with good justification. Insight linked to practice with good detail. | Visualizations are all relevant and clear, providing clear insights into the dataset. | All design choices are justified, with some references to Tufts Principles. |
| *Good (50 - 59%)* | Solid EDA conducted with rationale provided. Insights are mostly relevant but may lack depth. | Good explanation of most methods with adequate justification. Some links to insights but may lack detail. | Visualizations are mostly relevant and clear, providing useful insights into the dataset. | Most design choices are justified, with some references to Tufts Principles. |
| *Acceptable (40 - 49%)* | Basic EDA conducted, with minimal rationale. Insights are superficial or unclear. | Basic explanation of some methods used; justifications are weak or unclear. | Visualizations are present but may lack relevance or clarity; insights are limited. | Some design choices are mentioned but lack sufficient justification; few references to principles. |
| *Fail (> 39%)* | Inadequate EDA performed; little to no rationale or insight provided. | Little to no explanation of data preparation methods; no justifications provided. | Visualizations are missing or irrelevant, providing no insights into the dataset. | Poor or no justification of design choices; no references to Tufts Principles. |

## Grading Criteria Data Preparation & Visualisation

**The Irish Grading System**

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 experienced 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.

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 and review the marking criteria outlined in this assignment brief for a breakdown of the marking criteria for this specific assignment.

**Additional Information**

* Lecturers are not required to review draft assessment submissions. This may be offered at the lecturer’s discretion.
* In accordance with CCT policy, feedback to learners may be provided in written, audio or video format and can be provided as individual learner feedback, small group feedback or whole class feedback.
* Results and feedback will only be issued when assessments have been marked and moderated / reviewed by a second examiner.
* Additional feedback may be provided as individual, small group or whole class feedback. Lecturers are not obliged to respond to email requests for additional feedback where this is not the specified process or to respond to further requests for feedback following the additional feedback.
* Following receipt of feedback, where a student believes there has been an error in the marks or feedback received, they should avail of the recheck and review process and should not attempt to get a revised mark / feedback by directly approaching the lecturer. Lecturers are not authorised to amend published marks outside of the recheck and review process or the Board of Examiners process.
* Students are advised that disagreement with an academic judgement is not grounds for review.
* For additional support with academic writing and referencing students are advised to contact the CCT Library Service.
* For additional support with subject matter content students are advised to contact the [CCT Student Mentoring Academy](https://moodle.cct.ie/course/view.php?id=827)
* For additional support with IT subject content, students are advised to access the [CCT Support Hub](https://moodle.cct.ie/course/view.php?id=1861).