A picture containing icon

Description automatically generated

**CCT College Dublin Continuous Assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Programme Title:** | *MSc in Data Analytics* | | |
| **Cohort:** | *MSc in Data Analytics FT/SB+ (Feb24 start)* | | |
| **Module Title(s)**: | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* | | |
| **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)**: | *Sam Weiss*  *Taufique Ahmed/* *Marina Iantorno*  *Muhammad Iqbal*  *David McQuaid* | | |
| **Issue Date:** | 08/03/2024 | | |
| **Submission Deadline Date:** | 5/04/2024 | | |
| **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, GITHUB Link*  *Note that the maximum number of Jupyter Notebooks is 4* | | |
| **Feedback Method:** | **Results posted in Moodle gradebook** | | |
| **Feedback Date:** | *3 weeks after the last submission including PMC’s* | | |

**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**

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

**Statistics for Data Analytics**

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)

**Machine Learning for Data Analysis**

2. 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)

3. Implement a range of classification and regression techniques and detail / document their suitability for a variety of problem domains. (Linked to PLO 5)

4. Critically evaluate the performance of Machine Learning models, propose strategies to optimise performance. (Linked to PLO 3)

**Data Preparation & Visualisation**

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)

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

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

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

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 9 awards** |
| 90% + | Exceptional | 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 |
| 50 – 59% | Good | Attains all the minimum intended programme learning outcomes |
| 40 – 49% | Acceptable |
| 35 – 39% | Fail | 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 |

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 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.

**Acceptable and Unacceptable Use of AI**

|  |  |
| --- | --- |
| **Acceptable and Unacceptable Use of AI** | * 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   + Completing group work that your group has assigned to you   + Writing a draft of a writing assignment   + Writing entire sentences, paragraphs, papers, code fragments, functions, scripts 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. |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 

**Note ALL Students are required to use Git for any Assignments that they are working on.**

This means that ALL changes must be committed to Git 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 must also include ALL your lecturer's CCT emails as a collaborator on your account.

# Assessment Task

Students are advised to review and adhere to the submission requirements documented after the 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#/>

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. 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]

It is critical to be able to clearly and concisely describe the characteristics of the dataset being analysed. It provides an insight into the overall information contained within the dataset, the types of variables and the relationship between them. It is the first stepping stone within the overall analysis and often guides the subsequent steps (Kaur et al., 2018).

The Python pandas library was utilised to maximise efficiency and accuracy. The .types function was used to determine the data types that were contained in the data frame. The number of passengers, with column name VALUE, was numeric, as was the derived decimal year. The direction, type of movement and month were all categorical. There were 11 types of passenger movement recorded in the dataset. The categorical data was all nominal, with no variable being superior to another. The direction could be considered a dichotomous nominal variable as it had only two unique values, arriving to or departing from Ireland. The number of passengers entering/leaving the country was a discrete numerical value. Half of a person cannot travel and therefore this variable consisted of only whole numbers. The value could also be classified as a ratio variable as it has a definite zero point.

The dataset was obtained from data.gov.ie and was collected by the CSO. It is representative of the population who travelled from 1961 to 2010.

The measures of central tendency were taken for the dataset next. The .describe pandas function was used. This gave a mean value of 130228 with a standard deviation of 224501, indicating the data was skewed. The lower quartile, median and upper quartile values were 28723, 44200 and 124587 respectively. The mode does not apply to numeric data and did not reveal anything about the type of passenger movement as there was an equal number of entries for each category

A graph with a number of passengers

Description automatically generated

Figure AA: Histogram of passenger\_movement\_df

This indicated that the data was heavily skewed and this was visualised by plotting the values on a histogram. In Figure AA the data is observed as right skewed with the binwidth unspecified. A KDE was overlayed onto the plot to give an estimate for the probability density function. Histograms play a key role in determining the physical layout of continuous variables (Chen et al., 2008).

A graph of different colored squares

Description automatically generatedAn initial box plot was created containing all the data which showed there were a lot of outliers. For this reason, it was decided to take subsets of the dataset to create more meaningful plots, statistics and probability distributions. Boxplots are an important tool to highlight outliers in a dataset (Chen et al., 2008). Figure BB shows a side-by-side of the box plots for those arriving and departing by air and sea.

Outliers

Upper Quartile

IQR

Lower Quartile

Figure BB: Comparison of Boxplots

The data frame was filtered to show data pertaining to those who arrivedand departed by sea and air across the period. Taking a look at those arriving the two data frames (blue and yellow on the plot) have comparable medians of 177184 and 93685, denoted by the horizontal line in the box, it is clear that the two datasets are spread out over very different ranges. The whiskers on the boxplot are used to classify a point as being an outlier or not. They are calculated by multiplying the IQR by 1.5 (Chen et al., 2008). Values which lie outside of this are considered to be outliers and are denoted by a dot in Figure BB. The boxplot illustrates the variability in air compared to sea travel. While sea travel has remained relatively steady the aviation sector has increased significantly from 1961 to 2010. The IQR of the yellow plot is 40,590 and the blue is 117,161. This measure of spread indicates that there is much more variability within the air than sea travel sector. (add 5 number summary?)

A graph with many colored dots

Description automatically generated

Figure CC: Scatter plot from 1961 to 2010

The third chosen plot was a scatter plot, shown in figure CC. The year is plotted along the x-axis and the number of passengers moving in millions is along the y-axis. Up until the 1990s, the data remains relatively constant with a peak occurring in each year corresponding to the summer months followed by a trough in the winter. After 1990 air travel starts to increase considerably, both incoming and exiting the country. This reflects what was observed in the boxplot with the large variation in data for air travel. There is consistently more incoming passengers than departing for air travel, but the opposite is true for sea travel. Scatterplots are invaluable for showing the relationship between two variables and in the case of time data it can show trends over the given period. Air travel is clearly increasing while sea travel has remained relatively stable over the 50 year presented.

Boxplot - outliers

Scatter plot – trends over time

Histogram – frequency, skewed

1. 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. 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]

Probability distributions allow a numeric value to be assigned that an event will occur. Depending on the circumstances in question different distributions are appropriate. It is crucial to apply a relevant theory as the underlying assumptions will dictate the accuracy of the probability results. Probabilistic systems are based on the uncertain nature of events. They aim to interpret the range of possible outcomes, as opposed to the single outcome in deterministic systems given certain conditions, and compare the chances that various situations will arise based on collected data (Neil A. Weiss, 2012).

One such application of probability theory is the binomial distribution. It is a series of events or trials where either an event occurs or does not. A probability value Is assigned to both, the sum of which must be added to give 1. The events must be independent of one another, i.e. the probability of a passenger entering Ireland does not affect the probability of the next person entering.

There were three ways to enter the country, sea, air or cross-border travel, and since the binomial distribution requires only two outcomes a different question was posed.

The event which was chosen to explore was if a passenger travelling by sea was chosen at random what was the probability that the individual was arriving or leaving Ireland.

Binomial

Poisson – tending towards normal distribution (large data set)

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

Subset identified with approximate Gaussian shape.

1. 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]**

Data types

Removing unnecessary data

Replacing/dealing with missing data

Inbuilt pandas functions

Added columns (seasons)

1. You must also rationalise, justify, and detail all the methods used to prepare the data for ML (Scaling, Encoding, imputation etc…). **[0-40]**

Scaling – values varied by multiple magnitudes, min-max scaling used

Encoded the categorical data (season, method of transport, arriving or leaving)

1. Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. **[0-20]**

Visualisation notes

1. 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 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 modeling. **[0 - 20]**

Research one of those frameworks

CRISP-DM:

Cross-Industry-Standard-Process model for Data Mining (CRISP-DM) is a widely used data management framework in industry. Many data projects do not follow a specific framework but doing so can reduce costs and completion timelines, as well as promote best practice implementation. It has been applied to a wide range of data projects dealing with biological processes, software projects, cyber forensics and healthcare (Ayele, 2020). For this reason is has been described as industry independent. This indicates it has a high degree of flexibility, an advantageous attribute (Schröer et al., 2021).

It can be divided into 6 main processes: Business and data understanding, Data preparation, Modelling, Evaluation and Deployment. These pillars of the framework offer a data team structure and ensure that a data science team logically approaches the problem (Schröer et al., 2021).

The shortcomings of CRISP-DM were discussed by Studer et al. CRISP-DM is focused more on static analysis and is not set up to handle real-time data. This limits the dynamic decision making capabilities of this approach in a machine-learning setting. Furthermore, there is a gap in the framework relating to standards and quality assurance. As with all machine learning results the model must be assessed within their given conext (Studer et al., 2021).

1. 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]**

Linear Regression

1. Show the results of two or more ML modeling 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]**
2. 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]**

**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 explanations of your code choices (code quality standards should also be applied). **[0-50]**

Functions (encode, year conversion), for loops

1. 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. **[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.**

**CA1 NOTE DO NOT ZIP YOUR SUBMISSION FILES, ALL FILES MUST BE SUBMITTED INDIVIDUALLY**

# Submission Requirements

* All assessment submissions must meet the minimum requirements listed below. Failure to do so may have implications for the mark awarded.
* All assessment submissions must:
  + 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.
  + GITHUB Link
  + Be submitted by the deadline date specified or be subject to late submission penalties.
  + Be submitted via Moodle upload.
  + Use [Harvard(Ayele, 2020) 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.

**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 requested by *contacting the appropriate lecturer.* 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/mod/forum/view.php?id=55148)
* 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).

Ayele, W.Y., 2020. Adapting CRISP-DM for Idea Mining. IJACSA 11. https://doi.org/10.14569/IJACSA.2020.0110603

Chen, C., Härdle, W., Unwin, A., 2008. Handbook of Data Visualization. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-33037-0

Kaur, P., Stoltzfus, J., Yellapu, V., 2018. Descriptive statistics. Int J Acad Med 4, 60. https://doi.org/10.4103/IJAM.IJAM\_7\_18

Neil A. Weiss, 2012. Introductory STATISTICS, 9th ed. Pearson Education, Arizona State University.

Schröer, C., Kruse, F., Gómez, J.M., 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Studer, S., Bui, T.B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., Müller, K.-R., 2021. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. MAKE 3, 392–413. https://doi.org/10.3390/make3020020