­CCT College Dublin

|  |  |
| --- | --- |
| Module Titles: | Data Preparation & Visualisation, Machine Learning for Data Analytics, Programming for Data Analytics, Statistics for Data Analytics |
| Assessment Title: | Integrated CA, Ireland Tourism |
| Lecturer Names: | David McQuaid, Dr. Muhammad Iqbal, Sam Weiss, Taufique Ahmed |
| Student Full Name: | Ricardo Alves de Souza |
| Student Number: | 2018499 |
| Assessment Due Date: | 05/04/2024 |
| Date of Submission: | 05/04/2024 |

Declaration

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Title: Integrated CA, Ireland Tourism

Author: Ricardo Alves de Souza

E-mail: [2018499@student.cct.ie](mailto:2018499@student.cct.ie?subject=Integrated%20CA1)

GitHub Address: <https://github.com/ricardoasouz/2024---MSc-in-Data-Analytics---Feb---FT/tree/main/MSC_DA_CA1>

Table of Contents

**No table of contents entries found.**

**No table of figures entries found.**

# Abstract

It is a useful and common practice to put the abstract in Times New Roman 12-point italics. Throughout this document, the styles used reflect the styles we suggest you use in your scientific report.

# Introduction

Format your submission using the following styles as a guide. If you are completing your report for a specific assignment, make sure you follow any formatting guidelines provided by your supervisor, lecturer or tutor.

# Difference between Discrete Data and Continuous Data.

## Definition of discrete data.

Discrete data is a type of quantitative data that includes no divisible figures and statistics you can count.

You typically write discrete data points as numbers that represent exact values.

You can often describe discrete data by using the phrase "the number of," such as the number of customers in a store.

Discrete data usually represents single events that have already occurred.

When reviewing discrete data, you can analyse exact figures like units sold on a specific day or the hours an employee worked during a certain week.

## Definition of continuous data.

Continuous data is a type of quantitative data that represents precise measurements of nearly any numeric value.

Often, a continuous data variable has many decimal points because it's an exact measurement between two defined points.

Measuring this type of data is common in industries requiring specific data, such as health care, manufacturing and research and development.

Continuous data may change over time, allowing businesses to analyse their operations and predict future trends.

For example, a business may track the amount of time it takes a team to complete projects.

# Definition of Descriptive Statistics

Descriptive statistics summarize certain aspects of a data set, population or sample, using numeric calculations.

## Common descriptive statistics in general include:

* Count / Size
* Minimum
* Maximum
* Sum
* Mean
* Median
* Average
* Midrange
* Standard deviation
* Quartiles

But this not all, descriptive statistics can have more different types of analyses, like:

-

-

# Formulas and Calculations used on Descriptive Statistics

## Minimum

Ordering a data set:

x1 ≤ x2 ≤ x3 ≤ ... ≤ xn

The minimum is the smallest value x1, on the left, starting from lowest to highest.

Min = 𝑥1 = min(𝑥𝑖)𝑛𝑖 = 1

## Maximum

Ordering a data set:

x1 ≤ x2 ≤ x3 ≤ ... ≤ xn

The maximum is the largest value xn, on the right, starting from lowest to highest.

Max=𝑥𝑛=max(𝑥𝑖)𝑛𝑖=1

Range

The difference between the minimum and maximum.

Range=𝑥𝑛−𝑥1

## Sum

The total of all data values.

x1 + x2 + x3 + ... + xn

Sum=∑𝑖=1𝑛𝑥𝑖

## Size / Count

The number of data points on data set.

Size=𝑛=count(𝑥𝑖)𝑛𝑖=1

## Mean

The sum of all data divided by your size. Mean is also known as the “Average”.

For a Population

𝜇=∑𝑛𝑖=1𝑥𝑖𝑛

For a Sample

𝑥⎯⎯⎯=∑𝑛𝑖=1𝑥𝑖𝑛

## Median

Ordering a data set x1 ≤ x2 ≤ x3 ≤ ... ≤ xn from lowest to highest value,

the median is the numeric value separating the upper half of the ordered sample data from the lower half.

If n is odd the median is the center value. If n is even the median is the average of the 2 center values.

If n is odd the median is the value at position p where

𝑝=𝑛+12

𝑥˜=𝑥𝑝

If n is even the median is the average of the values at positions p and p + 1 where

𝑝=𝑛2

𝑥˜=𝑥𝑝+𝑥𝑝+12

- Mode -

The mode is the value or values that occur most frequently in the data set.

A data set can have more than one mode, and it can also have no mode.

- Standard Deviation -

Standard deviation is a measure of dispersion of data values from the mean.

The formula for standard deviation is the square root of the sum of squared differences from the mean divided by the size of the data set.

For a Population

𝜎=∑𝑛𝑖=1(𝑥𝑖−𝜇)2𝑛‾‾‾‾‾‾‾‾‾‾‾‾‾‾√

For a Sample

𝑠=∑𝑛𝑖=1(𝑥𝑖−𝑥⎯⎯⎯)2𝑛−1‾‾‾‾‾‾‾‾‾‾‾‾‾‾√

- Variance -

Variance measures the dispersion of data from the mean. The formula for variance is the sum of squared differences from the mean divided by the size of the data set.

For a Population

𝜎2=∑𝑛𝑖=1(𝑥𝑖−𝜇)2𝑛

For a Sample

𝑠2=∑𝑛𝑖=1(𝑥𝑖−𝑥⎯⎯⎯)2𝑛−1

- Midrange -

The midrange of a data set is the average of the minimum and maximum values.

MR=𝑥𝑚𝑖𝑛+𝑥𝑚𝑎𝑥2

- Quartiles -

Quartiles separate a data set into four sections.

The median is the second quartile Q2.

It divides the ordered data set into higher and lower halves.

The first quartile, Q1, is the median of the lower half not including Q2.

The third quartile, Q3, is the median of the higher half not including Q2.

This is one of several methods for calculating quartiles.[1]

- Interquartile Range -

The range from Q1 to Q3 is the interquartile range (IQR).

𝐼𝑄𝑅=𝑄3−𝑄1

- Outliers -

Potential outliers are values that lie above the Upper Fence or below the Lower Fence of the sample set.

Upper Fence=𝑄3+1.5×𝐼𝑄𝑅

Lower Fence=𝑄1−1.5×𝐼𝑄𝑅

- Sum of Squares -

The sum of squares is the sum of the squared differences between data values and the mean.

For a Population

𝑆𝑆=∑𝑖=1𝑛(𝑥𝑖−𝜇)2

For a Sample

𝑆𝑆=∑𝑖=1𝑛(𝑥𝑖−𝑥⎯⎯⎯)2

- Mean Absolute Deviation -

Mean absolute deviation[2] is the sum of the absolute value of the differences between data values and the mean, divided by the sample size.

For a Population

𝑀𝐴𝐷=∑𝑛𝑖=1|𝑥𝑖−𝜇|𝑛

For a Sample

𝑀𝐴𝐷=∑𝑛𝑖=1|𝑥𝑖−𝑥⎯⎯⎯|𝑛

- Root Mean Square -

The root mean square describes the magnitude of a set of numbers.

The formula for root mean square is the square root of the sum of the squared data values divided by n.

𝑅𝑀𝑆=∑𝑛𝑖=1𝑥2𝑖𝑛‾‾‾‾‾‾‾‾√

Standard Error of the Mean

The standard error of the mean is calculated as the standard deviation divided by the square root of the count n.

For a Population

𝑆𝐸𝜇=𝜎𝑛√

For a Sample

𝑆𝐸𝑥⎯⎯⎯⎯=𝑠𝑛√

- Skewness -

Skewness[3] describes how far to the left or right a data set distribution is distorted from a symmetrical bell curve.

A distribution with a long left tail is left-skewed, or negatively-skewed.

A distribution with a long right tail is right-skewed, or positively-skewed.

For a Population

𝛾1=∑𝑛𝑖=1(𝑥𝑖−𝜇)3𝑛𝜎3

For a Sample

𝛾1=𝑛(𝑛−1)(𝑛−2)∑𝑖=1𝑛(𝑥𝑖−𝑥⎯⎯⎯𝑠)3

- Kurtosis -

Kurtosis[3] describes the extremeness of the tails of a population distribution and is an indicator of data outliers.

High kurtosis means that a data set has tail data that is more extreme than a normal distribution.

Low kurtosis means the tail data is less extreme than a normal distribution.

For a Population

𝛽2=∑𝑛𝑖=1(𝑥𝑖−𝜇)4𝑛𝜎4

For a Sample

𝛽2=𝑛(𝑛+1)(𝑛−1)(𝑛−2)(𝑛−3)∑𝑖=1𝑛(𝑥𝑖−𝑥⎯⎯⎯𝑠)4

- Kurtosis Excess -

Excess kurtosis describes the height of the tails of a distribution rather than the extremity of the length of the tails.

Excess kurtosis means that the distribution has a high frequency of data outliers.

For a Population

𝛼4=∑𝑛𝑖=1(𝑥𝑖−𝜇)4𝑛𝜎4−3

For a Sample #(This is just Kurtosis in MS Excel and Google Sheets)

𝛼4=𝑛(𝑛+1)(𝑛−1)(𝑛−2)(𝑛−3)∑𝑖=1𝑛(𝑥𝑖−𝑥⎯⎯⎯𝑠)4−3(𝑛−1)2(𝑛−2)(𝑛−3)

- Coefficient of Variation -

The coefficient of variation describes the dispersion of data around the mean.

It is the ratio of the standard deviation to the mean.

The coefficient of variation is calculated as the standard deviation divided by the mean.

For a Population

𝐶𝑉=𝜎𝜇

For a Sample

𝐶𝑉=𝑠𝑥⎯⎯⎯

- Relative Standard Déviation -

Relative standard deviation describes the variance of a subset of data from the mean.

It is expressed as a percentage.

Relative standard deviation is calculated as the standard deviation times 100 divided by the mean.

For a Population

𝑅𝑆𝐷=[100×𝜎𝜇]%

For a Sample

𝑅𝑆𝐷=[100×𝑠𝑥⎯⎯⎯]%

- Frequency -

Frequency is the number of occurrences for each data value in the data set.

Frequency is used to find the mode of a data set.

### Normal Text

In general, use a 12-point Times New Roman font, or other Roman font with serifs.

### Abstract and Keywords

Every submission should begin with an abstract of about 100 words in the normal text style but italicized. The abstract should be a concise statement of the problem, approach, findings, and conclusions of the work described. Keywords will be taken from your submission form and added when the publication is assembled.

### References and Citations

In the body of the text, cite the references like so (author, year). Use the standard Harvard format for listing references at the end of your report. You can refer to the reference list at the end of this document as an example.

### Page Numbers and Footnotes

Make sure you number pages unless explicitly asked not to. (For example, many journals request that page numbers are *not* used in submission for publication.) In general, avoid the use of footnotes. Stick to using references and citations.

## Sections

Breaking your report into sections can make it much easier to read. Main sections (Introduction, Materials and Methods, Results and Discussion/Conclusions should generally be in Arial 12-point bold title-case or follow specific instructions given with your real assignments. The use of screenshots or diagrams to enhance your explanations is encouraged.

### Subsections

Arial 12-point title case.

#### Sub-subsection headings

Arial 12-point italic.

**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).

**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 libraries) to complete the analysis required. All of this is to be implemented in a Jupiter 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]**
2. In a dedicated section in your report, discuss your use of aspects of various programming paradigms in developing 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]**

**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)

**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 justified, and 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. 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]

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

5. Explain the importance of the distributions used in points 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]

**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, and propose strategies to optimise performance. (Linked to PLO 3)

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. Explain 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), and choose appropriate hyperparameters for the optimal outcomes of Machine Learning models using an approach of hyperparameter tuning, 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]**

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]**

**Data Preparation & Visualisation**

1. Discuss the concepts, techniques and processes underlying data visualisation to

critically evaluate visualisation approaches concerning 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 to improve the performance of machine learning models. (linked to PLO 2, PLO 4)

# 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 the 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]**

## Figures

Use Normal Times New Roman 12-point font for figure captions.

## Writing Style

* Use a simple writing style – if five words convey your meaning don’t use ten.
* Make use of headings and sub-headings to break the text and your key ideas into readable sections.
* Use past tense apart from descriptions of existing knowledge and results when you should use the present tense.
* Be specific about observations and quantities.
  + Avoid the use of adverbs (quickly, rapidly, slowly).
  + If a word you are using ends in 'ly' the chances are you can provide a measurement instead (over five minutes *instead of* slowly).
  + Be careful with adjectives which refer to quantities (few, lots, many).
  + Provide actual amounts (three Mayflies *instead of* a few Mayflies).
* Always back up your assertions with data or logical arguments and references.
* Use a spell checker or dictionary before you print out or submit your final report.
* If you are unsure about grammar ask a friend or someone good at writing to read through your draft report.
* Do not use derogatory language and be careful and considerate if you criticise the work of others. Stick to the arguments and facts.

## Acknowledgements

The idea and parts of the formatting for this template came from the Association for Learning Technology (UK) ALT-C 2004 Research Paper Format Template.

## References

Knisely, K. 2005, *A Student Handbook for Writing in Biology*, 2nd Edition, Smauer Freeman.

Pechenik, J. A. 1997, *A Short Guide to Writing about Biology*, 3rd Edition, Addison-Wesley.

Silyn-Roberts, H. 1996, *A Practical Handbook for Science, Engineering and Technology Students,* Longman.

Harvard Style *UseIt,* Available at http://www.library.uq.edu.au/training/citation/harvard.html