School of Arts, Humanities and Social Science

Module title and code: Mathematics for Data Science (CMP020L011A)

Title of coursework: Mathematics for Data Science Coursework

Learning outcomes:	1. Obtain and utilize appropriate Attribute from the given Dataset 2. Interpret Data correctly 3. Clean, Process and transform data where necessary
Assessment weighting	50%
Maximum mark	100
Submission details (e.g. submission link)	Moodle Submission
Word limit (if applicable)	1000 words
Date set	A dataset related to (i) HR information from a company is also given, along with a dataset showing data related to (ii) Covid-19 deaths, and also a dataset related to (iii) Asian American quality of life.
Deadline	8th Jan 2025 (Wednesday)
Feedback and marks	Please insert details (verbal or via rubric, etc.) and date
Assessment setter's name	Dr Michael Ng

Academic Misconduct:

"Academic integrity and honesty are fundamental to the academic work you produce at the University of Roehampton. You are expected to complete coursework which is your own and which is referenced appropriately. The university has in place measures to detect academic dishonesty in all its forms. If you are found to be cheating or attempting to gain an unfair advantage over other students in any way, this is considered academic misconduct, and you will be penalised accordingly."

Further details about "Student Code of Conduct" and "Disciplinary Regulations" can be found at:

https://www.roehampton.ac.uk/corporate-information/policies/

1. Assessment introduction

In this Coursework you are to solve a set of tasks related to ***Data processing, perform simple statistical operation(s), use probability distribution function.*** Throughout this process of completing the task, you must also provide discussion based on your understanding which should reflect that you have appropriate understanding about "Data" (in terms of attribute types, e.g., Qualitative & Quantitative, Discrete or Continuous and based on the correct interpretation about Data, if you have chosen the methodologies for performing operations correctly.

For this Coursework, the several datasets have been provided. You will need to work with several of the attributes to complete the requirements of this coursework. You can choose attributes from a single data set or multiple, it is your choice.

2. Description of the Datasets

The dataset provided to you is an open-source dataset that contains information about the employees and stakeholders who work/previously worked in a giant company. To protect the privacy and maintain the data protection ideologies, the name, address, date of birth and other types of personal information, direct contact number or, email address of the employees are not included in this Dataset. The information provided are the age, role, highest educational qualification, marital status, their account balance, along with those, if they own any housing on their own, if they have any loan taken against their job, the duration of their loans and if they have defaulted over their loan repayments. Also, the number of campaigns they have organized/taken part-in are given. You will also, see the day and month information of when they started their journey with the company is also given here.

A dataset related to (i) HR information from a company is also given, along with a dataset showing data related to (ii) Covid-19 deaths, and also a dataset related to (iii) Asian American quality of life.

It is up to you which attributes you pick from the given data sets to apply the various mathematical functions. For example, you can use marital status attribute to calculate binomial distribution.

3. Challenges in this Coursework

- Handling a considerably large dataset by performing statistical and mathematical operations on it.
- Selecting attributes from the dataset to satisfy self-defined criteria.

4. Task list

Task 1

- Selecting attribute(s) to perform the following ***"Measure of central tendency":***
- Mean, Median and Mode.

Hint: You can formulate scenarios like:

- The balance for a particular educational background of people, finding the "Mean" of their balance.
- The "Mode" (most occurring) marital status of the people who were blue collar, etc.

1. In your report, discuss about your selected attribute(s) for the purpose and rationale behind it within 200- 250 words) for Task-1.

2. You must provide screenshots of each operation/execution of the code (for mean, median, mode) created by you with proper commenting.

Task 2

- Selecting attribute(s) to perform the following ***"Measure of Spread/Dispersion":*** Range, Variance and Standard Deviation.

Hint: You can formulate scenarios like:

- The balance for a particular type of job (e.g., admin) and for that the measure of spread can be determined.

1. In your report, discuss about your selected attribute(s) for the purpose and rationale behind it within 200- 250 words) for Task-2.
2. You must provide screenshots of each operation/execution of the code (Range, Variance and Standard Deviation) created by you with proper commenting.

Task 3

- Selecting at least **TWO attributes** appropriately to construct ***confidence intervals***.

- In your report, discuss your selected attribute for the purpose and rationale behind it within 200- 250 words) for Task-3.

- You must provide screenshots of your codes' operation/execution with proper commenting.

Task 4

- Selecting at least **TWO attributes** appropriately to perform ***hypothesis testing***.

- In your report, discuss about your selected attribute for the purpose and rationale behind it within 200- 250 words) for Task-4.

- You are also, to provide the screenshot of your operation/execution with proper commenting.

Deliverables (what you will need to submit):

All the following must be submitted within the due date to be considered for a complete submission for this assessment.

(1)Report - your report should contain

The report should use 12-point font in Times New Roman, 1-inch margins, and double line spaced. The report should be properly paged, paragraphed, and sectioned, and include the following sections in order in a (.pdf) file.

I. Cover Page

II. Table of Content

III. Tasks:
Task:1 Discussion (max. 250 words) along with Necessary Screenshot(s) (excluded from word count)
Task:2 Discussion (max. 250 words) along with Necessary Screenshot(s) (excluded from word count)
Task:3 Discussion (max. 250 words) along with Necessary Screenshot(s) (excluded from word count)
Task:4 Discussion (max. 250 words) along with Necessary Screenshot(s) (excluded from word count)

IV. References:
Referencing any point in your discussion section(s) If you have used ideas from anywhere other than the lecture notes and tutorial examples (e.g., from a book, the Internet, or a fellow student) then include a reference showing where the code or ideas came from and label your code carefully to show which parts are your and which parts are borrowed.

(2) Interactive Python File

The supporting file you submit with your report should be an interactive python file (**.ipynb** format), contains codes you implemented with proper titles (in text blocks)and comments (in code blocks) and the file name should be "your name" with ipynb extension, like **"your name.ipynb".**

If you are unable to upload the "your name.ipynb" file directly through your submission box, you can alternatively, upload a "zipped" or, "rar" folder that contains your .ipynb file in it.

Additional Information (if required):

- You will need to review **Week-1,4, 8, 9, 10, and 11** Seminars and Lab activities to complete these tasks.

- Remember, you MUST perform **Task-1,2,3 and ,4** in **python** using appropriate libraries e.g., NumPy, SciPy, Pandas etc. However, for preliminary preprocessing steps (if necessary) you may use excel only to some extent.

Assessment Criteria (Grade Boundaries or Rubric):

Criteria	**Excellent**	**Satisfactory**	**Not Satisfactory**	**Not Attempted**
Understanding Data, The understanding and interpretation of data	Understanding about the data is clear and appropriate for all tasks	Understanding about the data is good but not discussed in sufficient detail in some tasks	Understanding and interpretation of data is somewhat clear, and/or is not discussed in sufficient detail for most of the tasks	No understanding is reflected
Appropriateness of Discussion about the Data and the chosen/proposed a method(s)	Understanding about chosen method is appropriate and discussed accurately in the report	Understanding about the chosen method is ok but could be discussed better	Understanding and discussion about the method is somewhat ok but is lacking sufficient detail	The choice of method is not appropriate and/ or, not discussed at all
Execution How smoothly does the Codes Execute- any error there?	Codes for all tasks executed correctly	At least 3 task solutions executed correctly	At least 2 task solutions executed correctly	No codes provided or, most have syntax error
Code Quality and Commenting structures of codes used during the implementation	Code Quality is appropriate and commented properly	Code Quality Needs work and commenting is missing in some places	Code solutions do not cover the entirety of the problem definition and/or, not commented appropriately	No codes provided or, code provided is very poorly formatted and no comment provided
Quality of Information and/or, supporting Files ability to provide details or supporting documents to support the report.	The supplied ipynb file contains all of the code and text blocks to appropriately implement and explain the solutions to the tasks	The ipynb file contains all of the coding solutions for the tasks but some text blocks are missing/explanation provided are not sufficient	The ipynb file does not have all of the coding solutions for the tasks and some/all text blocks are missing/explanation provided are not sufficient	An ipynb file was submitted but it did not open/did not have any code and/or text blocks or no ipynb file was submitted
Report Submission ability to provide a submission that meets the requirements given.	Correct files submitted only with no issues	Correct files submitted only but some issues with files provided	Correct files submitted but with unwanted files and possibly issues in files provided	Incorrect/no files were submitted.