**Deliverables**

|  |
| --- |
| This assessed work has three deliverables:   * **it Repository Link**, to be submitted via the link * **Two Programming Tasks**, to be completed as instructed in the assignment specification and submitted via GitLab |

Programming for Data Science -   
Assignment Specification

Portfolio of Programming Exercises and Development Process Report

# Information

This single coursework assignment will run during the Autumn Term, and involves three separate components:

|  |  |
| --- | --- |
| **Component** | **Assignment %** |
| Git Repository Link Task | 5 |
| Programming Task 1 | 55 |
| Programming Task 2 | 40 |
| **Total** | **100** |

Clear instructions and coding requirements are given below. This document also contains an Appendix section with additional illustrations of what is required for Programming Task 1. Please read this document carefully and ensure you have included all requested elements when submitting your work.

PLEASE NOTE: All submissions **MUST** be your own work.

* Students are **NOT** permitted to collude or submit anyone else’s work as their own (including any partial solutions found on the web).
* An analysis of submissions will be made across the cohort to identify any evidence of collusion and/or plagiarism during the marking phase.
* The following link provides more information about assessment offences and any associated penalties:

# Git Repository Link Task

* The repository should be used to commit all coursework development work including the .csv data files used for each Programming Task.
* The initial repository commit should be an appropriately edited README.md file (including a student ID, student name and MSc programme information).

# B. Programming Task 1

* This programming task focuses on using Python to calculate a set of Pearson Correlation Coefficients for a given dataset using built-in functions and data structures ONLY.
  + For Task 1, you **MUST NOT** import any Python library functions. This means you cannot use Python modules such as *math* or libraries such as Pandas or NumPy.
* To print a the Pearson Correlation Coefficient for a given pair of Python Lists, it would be very easy to use the ***pearsonr()*** function provided in the SciPy library. However, this programming task is designed to assess your coding abilities and by preventing you from using this function you are forced to gain a deeper understanding of how to complete that task. To do this, you will need to develop your own algorithm. Try typing “calculate Pearson Correlation Coefficient by hand” into your favourite search engine.
* There is a single data file available for use in this programming task. The file contains a record of US police killings for the year 2015.
  + The data file is called ***task1.csv***. This CSV file includes a header row with multiple named data values. This file is available in the Assignments section on Blackboard
* Students are expected to follow appropriate coding standards such as code commenting, consistent identifier naming, code readability, and appropriate use of data structures.
* You are expected to identify the strengths/weaknesses of your approach. For this programming task, you are expected to write a reflective report which focuses on the process taken to develop a solution to the task. Please reflect on your experiences rather than simply describing what you did. The report should:
  + include an explanation of how you approached the task.
  + identify any strengths/weaknesses of the approach used.
  + consider how the approach used could be improved.
  + suggest alternative approaches that could have been taken instead of the one you used.

## B.1. Requirements

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Requirement** | **Description** | **Marks Available** |
| **FR1** | Develop a function to find the arithmetic mean | The function should take a Python List as a parameter and return its arithmetic mean. You should use the following list to test your function: 85, 29, 35, 55, 82, 45, 42, 21, 42, 60, 56, 30, 72, 56, 37, 65, 29, 14, 66, 43, 23, 39, 81, 56, 74, 29, 22, 27, 14, 66, 55, 33, 31, 66, 63, 41, 30, 48, 68, 58, 51, 44, 66, 34, 20, 71, 59, 57, 43, 48. | **2** |
| **FR2** | Develop a function to read a single specified column of data from a CSV file | The function should accept two parameters: the data file name and a column number. The column number specifies which of the columns to read. It can range between 0 and n-1 (where n is the number of columns). The function should return two values: a List containing all the specified column’s data values and the column name. You should use the ***task1.csv*** data file to test your function but your function should also work for other CSV files. An illustration of this is given in Appendix 1. | **6** |
| **FR3** | Develop a function to read CSV data from a file into memory | The ***task1.csv*** data file contains multiple columns of data values. This function should accept a single parameter: the data file name. It should make use of the function developed in FR2 to read all the columns of data from the data file and add them to a Dictionary data structure. The Dictionary should contain one entry for each column in the CSV data file. An illustration of this is given in Appendix 2. | **6** |
| **FR4** | Develop a function to calculate the Pearson Correlation Coefficient for two lists of data | This function should calculate the Pearson Correlation Coefficient for two lists of data. You should make use of the function developed in FR1. The function should take two lists of data (of equal length) as parameters. The function should return the calculated coefficient value. | **10** |
| **FR5** | Develop a function to generate a set of Pearson Correlation Coefficients for a given data file | This function should make use the function developed in FR4 to generate a Pearson Correlation Coefficient for every pair of columns in the data read into memory in FR3. The function should return a list of tuples, each tuple containing the two column names and associated correlation coefficient value. An illustration of this is given in Appendix 3. | **10** |
| **FR6** | Develop a function to print a custom table | This function should output the Pearson Correlation Coefficient for a subset of the column pairs generated in FR5. The function should take three parameters: list of correlation coefficient tuples, border character to use and which columns to include. High marks will be given for good use of padding in the table cells to improve readability. An illustration of this is given in Appendix 4. | **8** |

## B.2. Deliverables

* There is single deliverable for this task:
  + A Jupyter Notebook file (in .ipynb format) containing a complete solution to Programming Task.
    - You must use the template provided[[1]](#footnote-0).
  + The Jupyter Notebook should also include a Development Process Report written using Markdown reflecting on the process taken to develop a solution to this task
    - The report should not exceed 500 words.

## B.3. Submission

* The Jupyter Notebook file (incl. Development Process report) should be submitted electronically by including it in the student’s coursework GitLab repository.

## B.4. Grading Criteria

* Marks are allocated as follows:
  + up to 42 marks for the Python code solution
    - Marks will be awarded for each completed requirement **according to the level of completion**.
  + up to 5 marks for adherence to good coding standards.
  + up to 8 marks for the Development Process Report
    - Marks will be awarded for appropriate use of technical language, critical reflection on development process and quality of engagement with the reflective process

# C. Programming Task 2

* This programming task focuses on using NumPy/SciPy, Pandas, and Matplotlib/Seaborn to combine, clean and analyse two datasets related to student performance.
* Two data files have been provided for this task. These data files provide some real data from the Open University.
  + The ***task2a.csv*** data file contains background information about 26746 students including gender, age, disability status and score.
  + The ***task2b.csv*** data file contains information about the number of click events made by 26074 students using the University’s Virtual Learning Environment (VLE) system.
  + The files are available in the Assignment section on Blackboard
* Students are expected to follow appropriate coding standards such as code commenting, consistent identifier naming, code readability, and appropriate use of data structures.
* You are expected to identify the strengths/weaknesses of your approach. For this programming task, you must write a reflective report which focuses on the process taken to develop a solution to the task. Please reflect on your experiences rather than simply describing what you did. The report should:
  + include an explanation of how you approached the task.
  + identify any strengths/weaknesses of the approach used.
  + Consider how the approach used could be improved.
  + suggest alternative approaches that could have been taken instead of the one you used.

## C.1. Requirements

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Requirement** | **Description** | **Marks Available** |
| **FR7** | Read CSV data from two files and merge it into a single Data Frame | For this task you should use the ***task2a.csv*** and ***task2b.csv*** data files | **2** |
| **FR8** | Clean the merged data | Remove all rows from the merged data that contain a missing value in any column. Remove the following unnecessary column: ***region***, ***final\_result*** and ***highest\_qualification*** | **3** |
| **FR9** | Filter out unnecessary rows | Remove all rows where ***click\_event*** is smaller than ***10***. | **2** |
| **FR10** | Investigate the effects of engagement on attainment | Use an appropriate visualisation tool (such as Matplotlib or Seaborn) to investigate if there is any relation between the engagement (click events) and the level of attainment (score). You must include an explanation of your findings to achieve good marks for this requirement. | **6** |
| **FR11** | Test the hypothesis that engagement has some effect on levels of attainment | Using an appropriate Python library, test if there is any statistically significant relation between engagement and attainment. You must include an explanation of your findings to achieve good marks for this requirement. | **4** |
| **FR12** | Investigate the effects of disability on levels of attainment | Use an appropriate visualisation tool (such as Matplotlib or Seaborn) to investigate if there is any effect on levels of attainment due to disability. You must include an explanation of your findings to achieve good marks for this requirement. | **6** |
| **FR13** | Test if there is any difference between the attainment of disabled and non-disabled students | Using an appropriate Python library, test if there is any statistically significant difference between disabled and non-disabled students. You must include an explanation of your findings to achieve good marks for this requirement. | **4** |

## C.2. Deliverables

* There is single deliverable for this task:
  + A Jupyter Notebook file (in .ipynb format) containing a complete solution to Programming Task.
    - You must use the template provided [[2]](#footnote-1).
  + The Jupyter Notebook should also include a Development Process Report written using Markdown reflecting on the process taken to develop a solution to this task
    - The report should not exceed 500 words.

## C.3. Submission

* The Jupyter Notebook file (incl. Development Process report) should be submitted electronically by including it in the student’s coursework GitLab repository.

## C.4. Grading Criteria

* Marks are allocated as follows:
  + up to 27 marks for the Python code solution
    - Marks will be awarded for each completed requirement **according to the level of completion**.
  + up to 5 marks for adherence to good coding standards.
  + up to 8 marks for the Development Process Report
    - Marks will be awarded for appropriate use of technical language, critical reflection on development process and quality of engagement with the reflective process

# Appendix 1 – Example Column Extraction

For the following illustration, you should assume that the column number parameter is equal to ***1*** for the data file. There are 9 columns in this file and so column number can range between 0 and 8. For this data, the function would return two values: ***“Glucose”*** and ***[148,85,183,89,137,116,78,115,197,125,110,168,139]***

# Table Description automatically generated with medium confidence

# Appendix 2 – In-Memory Data Structure

Using the file illustrated in Appendix 1, the Dictionary produced in FR3 should look something like the illustration below. However, you must ensure that your function can work for any CSV file with a similar structure (such as a file with 5 columns and 100 rows or with 20 columns and 1000 rows).

**{**

**"Pregnancies" : [6,1,8,1,0,5,3,10,2,8,4,10,10],**

**"Glucose" : [148,85,183,89,137,116,78,115,197,125,110,168,139],**

**"BloodPressure" : [72,66,64,66,40,74,50,0,70,96,92,74,80],**

**"SkinThickness" : [35,29,0,23,35,0,32,0,45,0,0,0,0],**

**"Insulin" : [0,0,0,94,168,0,88,0,543,0,0,0,0],**

**"BMI" : [33.6,26.6,23.3,28.1,43.1,25.6,31,35.3,30.5,0,37.6,38,27.1],**

**"DiabetesPedigreeFunction" : [0.627,0.351,0.672,0.167,2.288,0.201, 0.248,0.134,0.158,0.232,0.191,0.537,1.441],**

**"Age" : [50,31,32,21,33,30,26,29,53,54,30,34,57],**

**"Outcome" : [1,0,1,0,1,0,1,0,1,1,0,1,0]**

**}**

# Appendix 3 – Statistical data based on In-Memory Data Structure

Using the in-memory data structure illustrated in Appendix 2, the List of Tuples produced in FR5 should look something like the illustration below. The full data output is too large to include here and so only some of the data has been included to help illustrate what is required. Remember that different CSV data files will result in different data being stored. The data file you have been provided with does not include any of the data shown below. Don’t be tempted to simply copy the result below into your Jupyter Notebook.

**[**

**("Pregnancies", "Glucose", 0.337),**

**("Pregnancies", "BloodPressure", -0.0025),**

**("Pregnancies", "SkinThickness", -0.7481),**

**("Pregnancies", "Insulin", -0.4772),**

**("Pregnancies", "BMI", -0.2313),**

**("Pregnancies", "DiabetesPedigreeFunction", -0.0872),**

**("Pregnancies", "Age", 0.3428),**

**("Pregnancies", "Outcome", 0.0167),**

**("Glucose", "Pregnancies", 0.337),**

**("Glucose", "BloodPressure", 0.1429),**

**("Glucose", "SkinThickness", -0.0028),**

**("Glucose", "Insulin", 0.4304),**

**("Glucose", "BMI", 0.0584),**

**("Glucose", "DiabetesPedigreeFunction", 0.2192),**

**("Glucose", "Age", 0.5328),**

**("Glucose", "Outcome", 0.5465),**

**+++++++ More data would be included here ++++++++**

**("Outcome", "Pregnancies", 0.0167),**

**("Outcome", "Glucose", 0.5465),**

**("Outcome", "BloodPressure", 0.0755),**

**("Outcome", "SkinThickness", 0.3585),**

**("Outcome", "Insulin", 0.3355),**

**("Outcome", "BMI", -0.0768),**

**("Outcome", "DiabetesPedigreeFunction", 0.2185),**

**("Outcome", "Age", 0.314)**

**]**

# Appendix 4 – Output table for Statistics

Using the output from the function produced in FR5, the following table outputs a subset of the available columns (as defined by the function parameter) using the border character **\*** and padding within the cells to ensure the table is readable:

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**\* Glucose \* BloodPressure \* BMI \* Age \***

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**\* Glucose \* - \* 0.1429 \* 0.0584 \* 0.5328 \***

**\* BloodPressure \* 0.1429 \* - \* -0.4522 \* 0.4194 \***

**\* BMI \* 0.0584 \* -0.4522 \* - \* -0.3847 \***

**\* Age \* 0.5328 \* 0.4194 \* -0.3847 \* - \***

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

1. There is a Jupyter Notebook template available in the Assignment folder on Blackboard - ***UFCFVQ-15-M\_Programming\_Task 1\_Template.ipynb*** [↑](#footnote-ref-0)
2. There is a Jupyter Notebook template available in the Assignment folder on Blackboard - ***UFCFVQ-15-M\_Programming\_Task\_2\_Template.ipynb*** [↑](#footnote-ref-1)