**A logo for college computing

Description automatically generated**

**Assessment Cover Page**

|  |  |
| --- | --- |
| ***Student Full Name: Conor Kirwan*** |  |
| ***Student Number: 24130*** |  |
| ***Module Title: Programming for AI*** |  |
| ***Assessment Title: CA1*** |  |
| ***Assessment Due Date: 31/Oct/2024*** |  |
| ***Date of Submission: 31/Oct/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.

**Abstract**

**[NOTE: This section is designated for the abstract. Abstracts are not assigned page numbers and should precede the table of contents. If an abstract is unnecessary for your work, please delete this page.]**

**Attention: All notes must be removed from the document before submission!!**

**Contents**

[Question 1 4](#_Toc181020356)

[Question 2 4](#_Toc181020357)

[Question 3 4](#_Toc181020358)

[Question 4 4](#_Toc181020359)

[Question 5 4](#_Toc181020360)

[References 4](#_Toc181020361)

[Appendix 4](#_Toc181020362)

[**Assessment Outline** 5](#_Toc181020363)

[**Description of Assessment Task** 5](#_Toc181020364)

[**Screen Recording (10 marks):** 7](#_Toc181020365)

[You are required to submit a video, no longer than 7 minutes, demonstrating your Jupyter Notebook in action, with a step-by-step explanation of the code. 7](#_Toc181020366)

[**Assessment Requirements** 7](#_Toc181020367)

[**Learning Outcomes:** 8](#_Toc181020368)

# Introduction

# Question 1

1. The categorize\_columns() function takes a dataset (in the form of a dictionary) as input, where the keys are column names, and the values are lists of data for each column.
2. The function initializes two empty lists: num\_cols and cat\_cols to store the names of numeric and categorical columns, respectively.
3. It prints the column names of the input dataset.
4. For each column in the dataset:  
     
   * It initializes a flag is\_numeric and two empty lists: num\_values and cat\_values.
   * It iterates over the values in the column and appends them to the appropriate list (num\_values or cat\_values) based on their data type (integer or non-integer).
   * For each value, it prints whether the value is numeric or not.
   * After checking all values in the column, it evaluates the following cases:
     + If num\_values is not empty and cat\_values is empty, the column is considered numeric, and its name is appended to num\_cols. It prints that the column is numeric.
     + If cat\_values is not empty and num\_values is empty, the column is considered categorical, and its name is appended to cat\_cols. It prints that the column is categorical.
     + If both num\_values and cat\_values are not empty, the column is considered to contain a mixture of numeric and categorical values. In this case, the column name is appended to both num\_cols and cat\_cols. It prints that the column contains a mixture.
5. Finally, the function prints the lists of numeric and categorical column names and returns them.
6. Two sample datasets are created: dataset1 and dataset2.  
     
   * dataset1 contains a single column 'A' with integer values.
   * dataset2 contains a single column 'B' with string values.

The categorize\_columns() function is called separately for each dataset, storing the returned lists of numeric and categorical column names in separate variables (num\_cols1, cat\_cols1, num\_cols2, and cat\_cols2).

# Question 2

1. The code starts by importing the required libraries: pandas and numpy.
2. Four sample datasets (dataset1, dataset2, dataset3, and dataset4) are created from Question 1.
3. The concatenate\_and\_analyze() function is defined, which takes the four datasets as input.
4. Inside the function, each dataset is converted to a DataFrame using pd.DataFrame.from\_dict().
5. The Pandas operations section begins:
   * The four DataFrames are concatenated along columns using pd.concat(axis=1).
   * Duplicate rows are removed from the concatenated DataFrame using df\_concat.drop\_duplicates().
   * The number of rows and columns in the resulting DataFrame is printed using df\_concat.shape.
6. The NumPy operations section begins:
   * The numeric columns in the DataFrame are identified using df\_concat.select\_dtypes(include=[np.number]).
   * The numeric data is selected from the DataFrame and converted to a NumPy array using df\_concat[numeric\_cols].to\_numpy().
   * A comment is added to explain that the .T represents the transpose of the numeric data.
7. If there are any numeric columns:
   * The correlation matrix for the numeric columns is calculated using np.corrcoef(numeric\_data.T).
   * A comment is added to explain that np.corrcoef() is used to calculate the correlation matrix.
   * The pair of columns with the highest correlation coefficient is identified by iterating over the correlation matrix.
   * The names of the columns with the highest correlation and their correlation coefficient are printed.
8. If there are no numeric columns, a message is printed indicating that no numeric columns were found in the DataFrame.
9. Finally, the concatenate\_and\_analyze() function is called with the sample datasets.

# Question 3

* This algorithm demonstrates the use of NumPy for numerical computations and array manipulation.
* Taking the mean of a sliding window is a common operation in signal processing and time series analysis.
* Calculating the minimum and maximum values of the mean data can provide insights into the overall distribution and range of the data.
* The [value]max calculation involves finding the maximum absolute value, which can be useful in various applications, such as normalization or scaling of data.
* This task showcases the ability to perform element-wise operations on NumPy arrays, which is essential in many data processing and analysis tasks.

# Question 4

# Question 5

# Conclusion

# References

# Appendix

**A screenshot of a computer

Description automatically generated**

**[NOTE: For the table of contents to function properly, you must use the correct headings for all your chapters and subchapters.**

**Heading 1: This is the main heading and should be employed for the primary title or chapter. For example: CHAPTER 1.**

**Heading 2: Use Heading 2 as a subheading. For instance: Chapter 1.1.**

**Heading 3: Heading 3 provides a more detailed breakdown, such as Chapter 1.1.1.**

**By adhering to this hierarchical structure, you ensure an organized and effective document outline, enhancing readability and navigation. However, you are not forced to use all 3 headings, usually heading 1 and 2 are sufficient.**

**The remainder of your text should be written using a normal font.]**

**References**

**Assessment Outline**

**Description of Assessment Task**

Complete the following tasks involving Python, NumPy, pandas, SQL, and exploratory data analysis (EDA). Your solutions should include user-defined functions, database integration, and an EDA coding and report on the Iris dataset. Submit a 1500-word report interpreting your findings, discussing the significance of each task in artificial intelligence.

**Q1 (15 marks)**

Select two datasets and create a user defined function in Python that leverages the fundamental features of the Python without using Pandas built in functions such as

(numeric\_df =df.select\_dtypes(include=['number'])

categorical\_df = df.select\_dtypes(exclude=['number'])

Your function should accept a Data Frame as input and categorize its columns into numeric and categorical types, then display the lists of numeric and categorical columns. Include a section in your report where you discuss your interpretation of this task and its significance within the field of artificial intelligence and make sure to run the code and attach a screenshot of your machine in the appendix of your report.

**Q2 (15 marks)**

Write a function, the function should achieve the following objectives:

1. In pandas:
   * Concatenate the two datasets, which you have used in Question1, along the rows.
   * Remove any duplicate rows.
   * Print the number of rows and columns in the resulting DataFrame.
2. In NumPy:
   * Calculate the correlation matrix for all numeric columns.
   * Identify the pair of columns with the highest correlation coefficient.
   * Print the names of these columns along with their correlation coefficient.

Include a section in your report where you discuss your comprehension of this task and its relevance in the field of data pre-processing and analysis using pandas and NumPy libraries. Make sure to run the code and attach a screenshot of your machine in the appendix of your report.

**Q3 (10 marks)**

Writ a python program to implement the below algorithm:

* Create a NumPy array with 1000 random elements (numbers) and take the mean of every 5 sample window.
* [datamean] < -- mean ( [ numpy\_array (1 : 5 : end ) ] )
* [data]min < -- min ( [ datamean ] )
* [data]max < -- max ( [ datamean ] )
* [ value ] max < -- max ( abs ( [data]max), abs([ [data]min])

Include a section in your report where you discuss your interpretation and significance of this task and make sure to run the code and attach a screenshot of your machine in the appendix of your report.

**Q4 (10 marks)**

Write a code for database integration and manipulating the data using SQL clause. Explain the significance of RDBMS such as SQL in the field of artificial intelligence (AI). Discuss why they are essential for tasks such as data storage, retrieval, pre-processing, and integration in AI applications. Describe the process of connecting to a MySQL database, creating tables, and storing values. Highlight how proficient utilization of databases and SQL enhances the efficiency and effectiveness of AI systems. Make sure to run the code and attach a screenshot of your machine in the appendix of your report.

**Q5 (15 marks)**

Conduct an exploratory data analysis (EDA) on the Iris dataset from scikit learn using Python. The objective is to provide a comprehensive overview of the dataset, including an examination of its features and target variables, as well as understanding the data structure—specifically its shape, data types, and any missing values. The analysis should include descriptive statistics for each feature, along with visualizations such as histograms and boxplots to explore feature distributions and relationships among the species. Additionally, a correlation matrix should be generated to identify significant correlations between features. Finally, the report should include interpretations and conclusions drawn from the analyses and visualizations, effectively summarizing insights gained from the EDA process. Make sure to run the code and attach a screenshot of your machine in the appendix of your report.

**Report (25 marks)**

A report (up to 1500 words excluding appendix section and conclusion) that provides a critical analysis and interpretation of the findings derived from all the previous questions.

**Screen Recording (10 marks):**

You are required to submit a video, no longer than 7 minutes, demonstrating your Jupyter Notebook in action, with a step-by-step explanation of the code.

**Assessment Requirements**

All assessment submissions must meet the following minimum requirements:

* Be submitted in the format outlined in the assignment summary table.
* Report (maximum 1500 words), and screen recorded video with explanation of code.
* You are required to create a GitHub account, upload your CA work, and grant access to my email address. Ensure to include the GitHub account link in your CA 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:

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:

1. Demonstrate knowledge of fundamental programming concepts.

(Linked to PLO 1)

1. Develop a requirements specification to Integrate file handling and database integration as part of given cognitive system proposal. (help / chat bot, game of life ) (Linked to PLO 4, PLO 5)