A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Thais Cristina Glauzer da Silva |
| *Student Number* | sba24101 |
| *Module Title* | Programming AI Concepts |
| *Assessment Title* | CA1 |
| *Assessment Due Date* | 31st October 2024 |
| *Date of Submission* | 29th October 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

This document is part of the Programming AI Concepts of the Higher Diploma in AI Applications (HCI) at CCT Dublin. The goal is to answer 5 questions proposed by our lecturer to showcase our learnings in the field of AI applications, namely Python and SQL for this first CA. This report contains screenshots from the Jupyter notebook used and reflections on each question and their importance for the AI field.

Link to Github containing the Jupyter notebook: <https://github.com/tg9292/CA1-Programming-for-AI>

Contents

[Q1 1](#_Toc181120398)

[Q2 1](#_Toc181120399)

[Q3 1](#_Toc181120400)

[Q4 2](#_Toc181120401)

[Q5 2](#_Toc181120402)

[References 3](#_Toc181120403)

[Appendix 4](#_Toc181120404)

# Q1

This exercise is focused on finding the basic information about your dataset. It’s paramount to understand the way your data is structured and the format of the data. In the field of AI this is even so more relevant as machine learning models are not capable of using categorical data. Understanding if the datasets used contain categorical data is paramount to proceeding with your analysis. In case categorical data is to be found, there’s a need to convert it to numerical data to use it in machine learning.

Another layer of this exercise is the request not to use Pandas. Pandas is broadly used in python as its functions enable quicker and intuitive analysis. By answering the problem without Pandas, we prove that python is sufficient for your data analysis needs, however, libraries have a huge potential in making the code shorter and intuitive. This question could have been answered by using the DataFrame.dtypes property in pandas.

# Q2

This question touches a key element in data analysis and AI. Combining different dataset is a powerful tool when feeding the machine and guaranteeing your dataset has enough diversity of sources when the conditions allow (datasets with similar structure or/and data that help you supporting your goals in a non-biased way with a balanced sample). Making sure you are removing duplicates guarantees a healthier population distribution in your dataset.

The steps suggested under the NumPy topic already help the analysts looking for insights in the dataset, saving them a lot of time. By identifying correlation coefficients automatically, your research can take a different turn, or maybe other correlations are discovered when previously neglected. The insights generated by running a simple correlation calculation can be of great help in solving problems.

In this example two completely different dataset were analysed together. It makes sense that the function was not able to correlate columns from one dataset (iris) with another (heart records). It would be surprising if they had any correlation, but let’s leave this for conspiratorial theorists.

In the first question (Q1) we were asked not to use the Pandas library and, on this exercise (Q2), we are told to apply Pandas and NumPy libraries. “Pandas is most commonly used for data wrangling and data manipulation purposes, and NumPy objects are primarily used to create arrays or matrices that can be applied to DL or ML models. Whereas Pandas is used for creating heterogenous, two-dimensional data objects, NumPy makes N-dimensional homogeneous objects.” (Ginsberg, 2024). This showcases that the analyst can resort to different extensions and libraries to solve different problems in python, therefore knowing your goals and the type of data you are handling is also key to understanding what to import in your analysis.

# Q3

In Q3 we are using random numbers instead of a dataset. This is a good exercise to help understand key mathematical elements and ideas in data analysis and statistics. The random.seed set by 42 guarantees that the same random numbers will be selected by anyone running the code, which enables multiple users to make the same analysis.

Mean calculations are excellent to understand trends in a population. The core idea is to remove extremes from your dataset, making its result more faithful to reality than a simple average calculation.

The minimum and maximum figures are essential to understanding the range of the data and the limits of the information collected. This interval sets the boundaries of the data collected and help understand fluctuations in future data collections. When it comes to machine learning, it’s very likely that your model will not suggest results beyond the minimum and maximum values and the tests should return most of the outputs closer to the mean value. Having outputs too far from these boundaries suggests anomalies in your algorithm or in the inputs.

# Q4

Relational Database Management Systems (RDBS) are very efficient when querying datasets. SQL offers an accessible language that enables easy deletion and creation of data, data integration (JOIN), filtering (SELECT, WHERE), aggregation (GROUP BY), cleaning (ISNULL) and much more, empowering analysts with a quick access to big data, making it crucial for AI applications.

Using SQL in AI systems enables collaboration between users and supports an easier way to clean and query the overwhelming amount of data needed for machine learning in AI. On the example above we demonstrate how to connect Python to an SQL server, how to create a table and feed it. Being able to simply add data to your huge database through a quick connection with SQL makes the handling of datasets for AI much more efficient than opening your big data and adding information directly there.

# Q5

This exercise combined basic understanding of the dataset with a deeper understanding driven from data visualization and correlation operations. Combined these can be powerful analytical tools, helping analysts and decision makers to understand problems and gather insights.

One important step taken was to analyse missing values, or null data, that could impact our dataset. We’ve found that the dataset contains no missing values, which we also demonstrated with a heatmap visual representation.

In our analysis, the boxplot chart was fundamental in the decision of what variables to choose from to analyse. It’s clear that the 3 species of iris vary more in petal length than any other metric collected. That is the value that probably will help us differentiate species from species more easily. This is the reason why we decided to take a deeper look into the petal length distribution per species on the bar chart. We can see that the setosa species usually has the longest petals. The distribution of petal length can also be observed in the histograms, where the chart has a similar pattern to the bar chart. This means that setosa is the biggest contributor in the big range found under petal length.

# References

[Chia, A. (2024).](https://www.datacamp.com/tutorial/python-boxplots?utm_source=google&utm_medium=paid_search&utm_campaignid=19589720824&utm_adgroupid=157098107175&utm_device=c&utm_keyword=&utm_matchtype=&utm_network=g&utm_adpostion=&utm_creative=716160943810&utm_targetid=dsa-2363008446900&utm_loc_interest_ms=&utm_loc_physical_ms=9209642&utm_content=&utm_campaign=230119_1-sea~dsa~tofu_2-b2c_3-row-p2_4-prc_5-na_6-na_7-le_8-pdsh-go_9-nb-e_10-na_11-na-oct24&gad_source=1&gclid=Cj0KCQjw7Py4BhCbARIsAMMx-_LNoe339OUW-0UoYRYLXFVYpEumhd8LRgB_g7gaByl1p5pH7svQH14aAreqEALw_wcB) “Python Boxplots: A Comprehensive Guide for Beginners”. *Datacamp*. Available at: <https://www.datacamp.com/tutorial/python-boxplots?utm_source=google&utm_medium=paid_search&utm_campaignid=19589720824&utm_adgroupid=157098107175&utm_device=c&utm_keyword=&utm_matchtype=&utm_network=g&utm_adpostion=&utm_creative=716160943810&utm_targetid=dsa-2363008446900&utm_loc_interest_ms=&utm_loc_physical_ms=9209642&utm_content=&utm_campaign=230119_1-sea~dsa~tofu_2-b2c_3-row-p2_4-prc_5-na_6-na_7-le_8-pdsh-go_9-nb-e_10-na_11-na-oct24&gad_source=1&gclid=Cj0KCQjw7Py4BhCbARIsAMMx-_LNoe339OUW-0UoYRYLXFVYpEumhd8LRgB_g7gaByl1p5pH7svQH14aAreqEALw_wcB>. Accessed on: 28th October 2024.

Dimond, R (2024). “Data Wrangling in Python”. *Social Science Computing Cooperative*. University of Wisconsin. Available at: <https://sscc.wisc.edu/sscc/pubs/dwp/>. Accessed on: 28th October 2024.

[Diop. D. (2023). “Unlock the Power of Python: A Comprehensive Guide to Unlock its Full Potential”. *Medium*. Available at: https://medium.com/@ditechdiop/unlock-the-power-of-python-a-comprehensive-guide-to-unlock-its-full-potential-66acdf716b70](file:///C:\Users\thais\Downloads\Diop.%20D.%20(2023).%20). Accessed on: 28th October 2024.

Ginsberg, C (2024). “Pandas vs. NumPy: Which Python Library is Better for Data Analytics?”. *Noble Desktop*. Available at: <https://www.nobledesktop.com/classes-near-me/blog/pandas-vs-numpy-for-data-analytics#:~:text=Pandas%20is%20most%20commonly%20used,makes%20N%2Ddimensional%20homogeneous%20objects>. Accessed on: 28th October 2024.

Google Cloud. “What is a relational database?”. Available at: <https://cloud.google.com/learn/what-is-a-relational-database?hl=en>. Accessed on: 28th October 2024.

Johansen, A. (2016). *Python: The Ultimate Beginner’s Guide*. Leipzig, Germany: Amazon Distribution GmbH.

SQream (2022). “The Role of SQL in Machine Learning”. Available at: <https://sqream.com/blog/the-role-of-sql-in-machine-learning/>. Accessed on: 28th October 2024.

# Appendix



















