Report AI Programming CA

**Student full name : Barbara Weltson Leite da Silva**

**Student Number:sba24096**

[**Question 1 1**](#_qw0jdnxz7vef)

[**Question 2 1**](#_9ndhojspiaif)

[**Question 3 2**](#_go1pywap7vqa)

[**Question 4 2**](#_hagqksuy0eol)

[**Question 5 3**](#_3x0jljq526vq)

[**Reference 4**](#_z2ko2654gu5x)

[**Appendix: 5**](#_bzifr7k0oaky)

# Question 1

When developing a program for AI, it's important to understand the differences between types of data, especially between numerical and categorical data. Numerical data is any data we can represent with numbers, like age, height, or temperature. Categorical data, on the other hand, represents groups or categories, such as colors, names, or types, where numbers cannot represent this data

Identifying the type of data in each column of our dataset helps us choose the right algorithms for building our model. Some algorithms work well with numerical data but cannot directly handle categorical data. By knowing which columns are numerical and which are categorical, we can decide how best to prepare the data for the model and select the algorithm that will give us the most accurate results.

# Question 2

When building an AI program, we also need to make sure that the datasets we use are easy to understand, clean, and accurate, especially when combining different datasets. To achieve this, we remove any duplicate entries to avoid errors that could impact our model’s accuracy. Here, we use the pandas library, which helps with data organization and cleaning, making it easier to work with and analyze the data.

In addition to pandas, we also use numpy, a foundational library in Python for scientific computing. With numpy, we can calculate the correlation matrix, which helps us see relationships between different variables. This is useful for identifying patterns and understanding which variables are connected or have similar behaviors.

Just as in Question 1, we combine (or concatenate) datasets and remove duplicates to make sure the data is ready to be analyzed. This preparation step helps ensure that our results are more reliable. By calculating the correlation matrix, we gain insights into the data’s structure, which allows us to spot patterns and make more informed choices in our analysis.

# Question 3

I’m not entirely sure of the significance of this specific task, but it seems like we are using a large array to represent a dataset in a simple way. By working with this array of random numbers, we can practice analyzing data that can also be random, which is something we often have to do in many real-life situations.

The task focuses on finding the mean, minimum, and maximum values in the data. Calculating the mean gives us an idea of the average value, which helps us understand what’s typical in our dataset

Finding the minimum and maximum values is also important because it helps us see the extremes in our data. These extreme values can highlight unusual or important points that may need our attention. So overall by understanding the mean, minimum, and maximum, we can make better decisions and spot trends that are important in various areas, like business, healthcare, or scientific studies. This kind of analysis is useful because with that, we can learn more from the data we have.

# Question 4

To create a successful AI program, it’s usually essential to feed a large amount of data into the model. This is where connecting to a database to store that data becomes very important. Databases act as a central place for managing data, making sure everything is organized and easy to find.

SQL (Structured Query Language) is key to this process because it lets us quickly pull data directly from its source. This quick access to data is crucial for training the AI system, as the model needs to handle large datasets efficiently to learn well. When we can get data fast, it speeds up the training process, allowing for quicker updates and improvements in how the model performs.

Having a well-organized database also helps the model train faster since it provides a clear way for the AI to find the information it needs.

Also, databases make it easier to combine different sets of data, giving us a full picture of the information available. This overall view is important for making smart decisions when training and evaluating AI models. SQL allows us to do things like join tables and merge datasets, which makes the process of bringing together information from various sources simpler. These features not only make data integration simpler but also ensure that the AI model has access to all the relevant information, leading to better predictions and insights.

Overall, using SQL databases in AI projects improves how we store and access data, which boosts the overall efficiency and effectiveness of AI systems. By taking advantage of these organized databases and their powerful querying features, developers can make the training process smoother and get the most out of their data.

# Question 5

The Iris dataset includes measurements for 150 flowers across three species: Setosa, Versicolor, and Virginica. These measurements cover things like petal and sepal sizes.

Some of the findings are as below:

* Setosa flowers stand out with shorter petals compared to the other species, which makes them easier to identify.
* Petal measurements (length and width) are very helpful when telling the species apart, while sepal measurements overlap more across species.
* Petal length and width are strongly related, especially for Versicolor and Virginica. These patterns hint that petal size will be a good predictor of species.

## Reference

Ali, M. (2020). *Handling Machine Learning Categorical Data with Python Tutorial*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/categorical-data?dc\_referrer=https%3A%2F%2Fmoodle.cct.ie%2F [Accessed 2 Nov. 2024].

NumPy (2022). *What is NumPy? — NumPy v1.19 Manual*. [online] numpy.org. Available at: https://numpy.org/doc/stable/user/whatisnumpy.html.

Pandas (2024). *pandas documentation — pandas 1.0.1 documentation*. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/docs/.

Wagavkar, S. (2023). *Introduction to The Correlation Matrix | Built In*. [online] builtin.com. Available at: <https://builtin.com/data-science/correlation-matrix>.

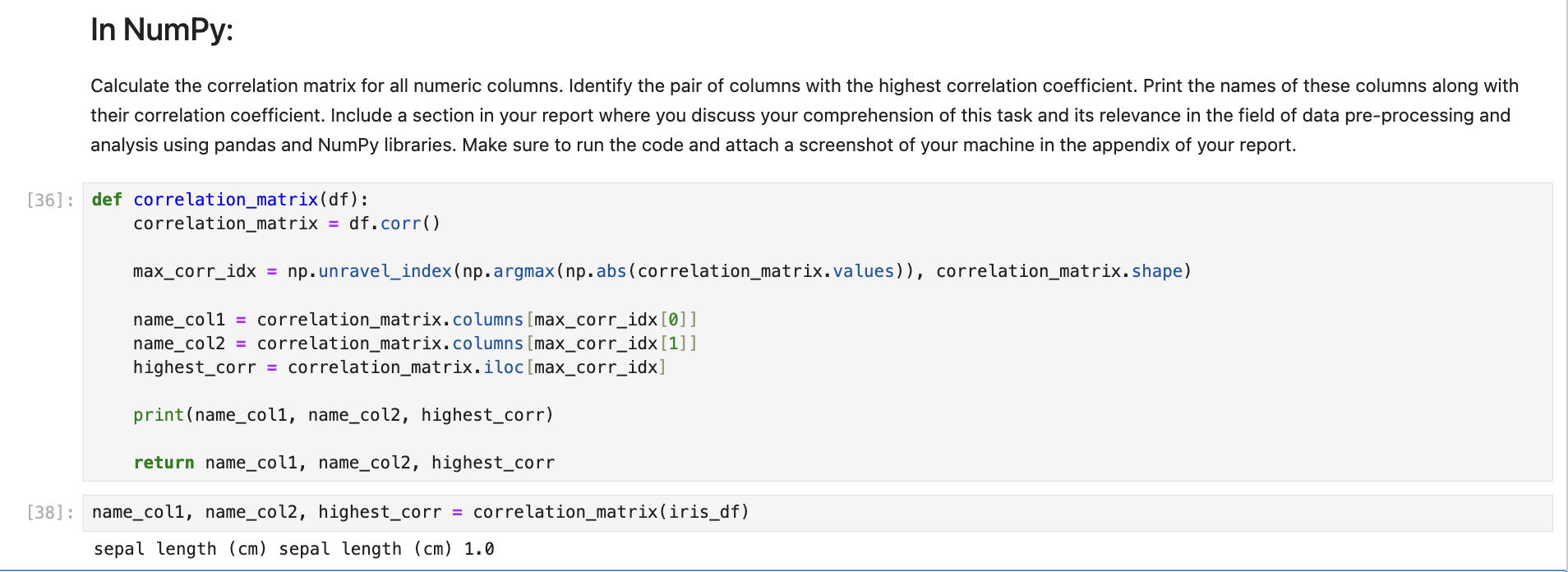
kaggle.com. (n.d.). *Intro to Exploratory data analysis (EDA) in Python*. [online] Available at: https://www.kaggle.com/code/imoore/intro-to-exploratory-data-analysis-eda-in-python.

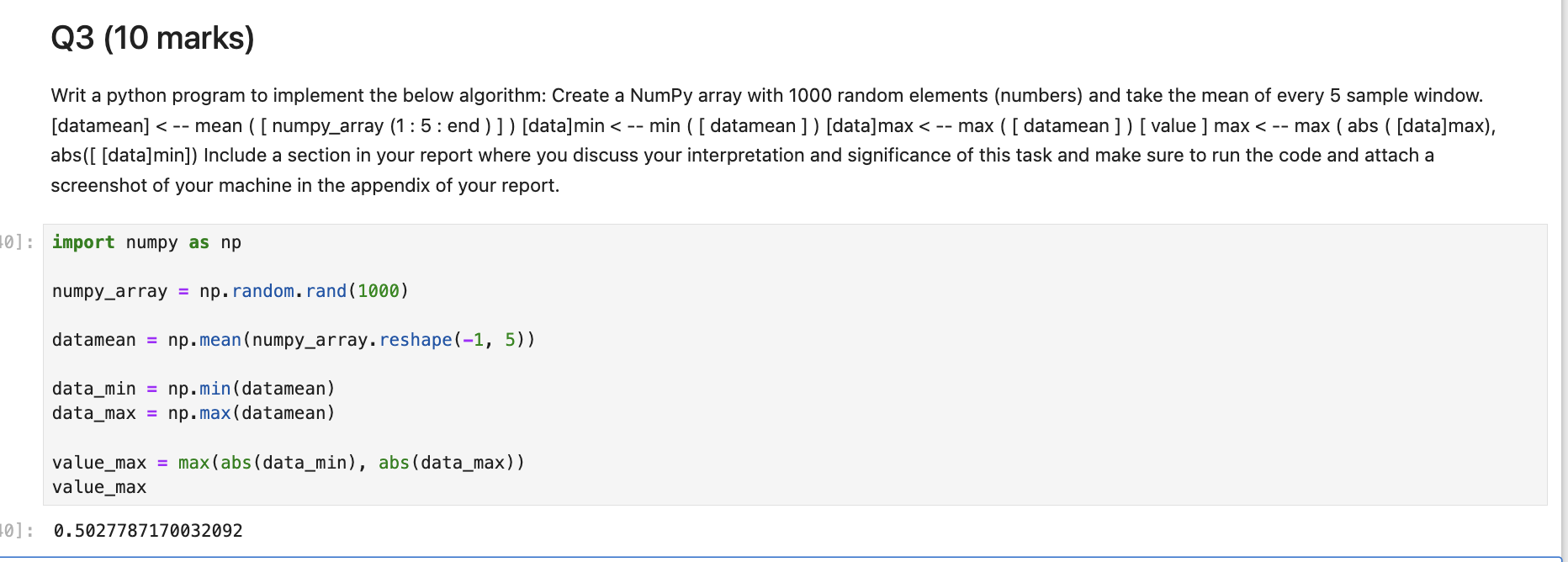
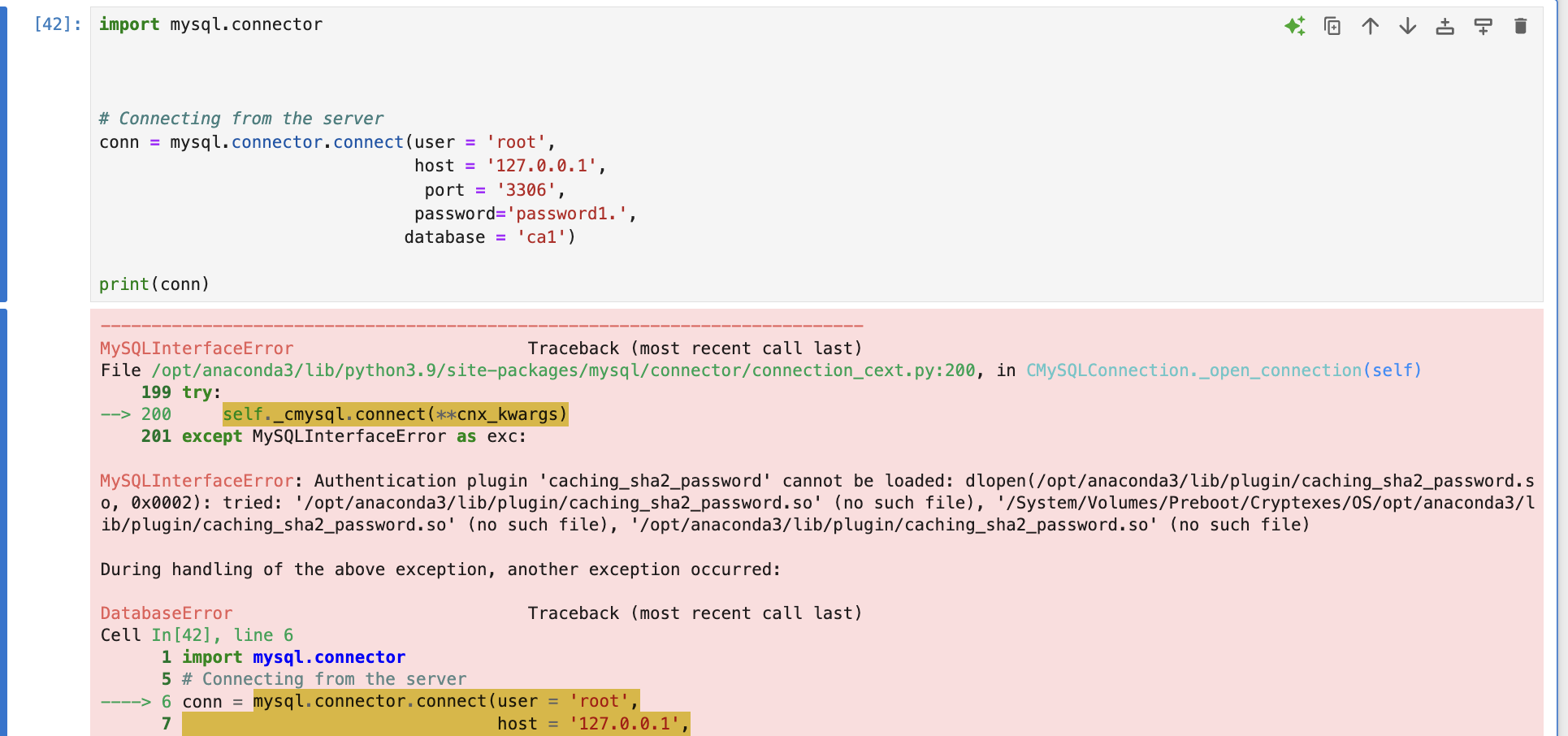
# Recording link <https://capture.dropbox.com/ZLzTug83oGU6XO7z>

## Appendix:







  
  
For question 4, I unfortunately keep getting errors but the code is there for the Database:  
  
  
  
Question 5:  