A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Santhosh Shanmugasundaram |
| *Student Number* | Sba24100 |
| *Module Title* | Programming for AI |
| *Assessment Title* | CA1 |
| *Assessment Due Date* | 31st October 2024 |
| *Date of Submission* | 31st 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.

Contents

[1. Categorize Numerical and Categorical Columns in Data Frame 1](#_Toc181308126)

[2. Concatenate Two Dataset and Calculate Correlation Matrix using NumPy 1](#_Toc181308127)

[2.1 Concatenate Two Dataset 1](#_Toc181308128)

[2.2 Calculate Correlation Matrix Using NumPy 2](#_Toc181308129)

[3. Find Mean, Min, Max of 1000 random numbers using NumPy Array 3](#_Toc181308130)

[4. Connect to MySQL DB, Create Table and Maintain data. 5](#_Toc181308131)

[4.1 Connect to MySQL Server 5](#_Toc181308132)

[4.2 Database and Tables 5](#_Toc181308133)

[4.3 Insert and Select Records 6](#_Toc181308134)

[5. Exploratory Data Analysis (EDA) on Iris Dataset 7](#_Toc181308135)

[6. References 10](#_Toc181308136)

[7. GitHub Link 11](#_Toc181308137)

# Categorize Numerical and Categorical Columns in Data Frame

**Scope** is to create a python function that takes data frame as input, categorizes the columns into numerical and categorical types and returns 2 results, one for numerical type and one for categorical type.

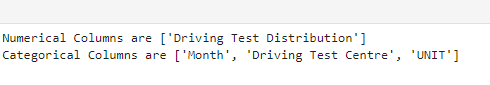
**Solution**:

A custom function ‘find\_type’ has been developed. This function will take 1 Data frame as input. Using ‘for’ statement, loops through all columns and finds the data type.

If the data type of the column is float64 or int64, then it will be marked as numerical type. Else, it will be marked as categorical.

For testing purpose, I have taken 2 datasets from ‘data.cso.ie’.  
1st dataset contains details about Ireland Driving Tester availability in each centre.

2nd dataset contains details about waiting time for each application applied for vehicle driving test at Ireland.





# Concatenate Two Dataset and Calculate Correlation Matrix using NumPy

## Concatenate Two Dataset

**Scope** is to create a python function that takes 2 data frame as input and concatenates the data in row.

**Solution**:

A custom function ‘concat\_df’has been developed. This function will take 2 Data frame as input and using the panda function ‘concat’, two data frames will be concatenated, and result is returned back as 1 data frame.

For testing, I have utilized the 2 data set used in question 1 and result has been shared below.

~~~~

Data is very important in building any model or data science project. When dealing with data, it is very common that data is coming from multiple sources in multiple formats. To arrive at a solution or even to process a model, it is important to have the data grouped together as one.  
For this purpose, it is very important to concatenate the data together.

## Calculate Correlation Matrix Using NumPy

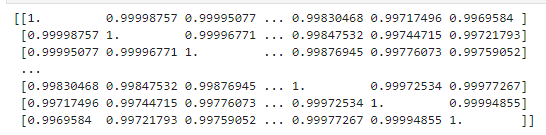
**Scope** is to create a correlation matrix for a sample data set and find pair of columns with highest correlation coefficient.

**Solution**:

The dataset of word happiness report 2023 has been taken from Kaggle and same as been downloaded in a csv file.  
Using the panda ‘read\_csv’ function, data has been loaded to a data frame.

To derive correlation matrix, it is important to use only the numeric data. Therefore, all the numeric columns from the dataset are passed to the NumPy function ‘corrcoef’.

The resultant has been displayed below.



To find the pair with highest correlation coefficient, we need to pass each set of two columns to the NumPy function ‘corrcoef’ and compare the result with each other.   
At the end, we have the highest correlation coefficient and the pair name as below,



Correlation is a key statistical concept that researchers employ to analyse connections within their data. It helps us to Understand the Relationship between Variables.

It is important for machine learning engineers to understand the correlation between variables in their models for following key reasons - Feature Selection, Reduce Bias, Multicollinearity, Interpretability and debugging.

By analysing the correlation between input features and output targets, researchers gain insights into which variables have the strongest impact on the model’s decisions. This helps ensure the model is actually learning meaningful patterns in the data rather than spurious correlations.   
(Ashraf)

# Find Mean, Min, Max of 1000 random numbers using NumPy Array

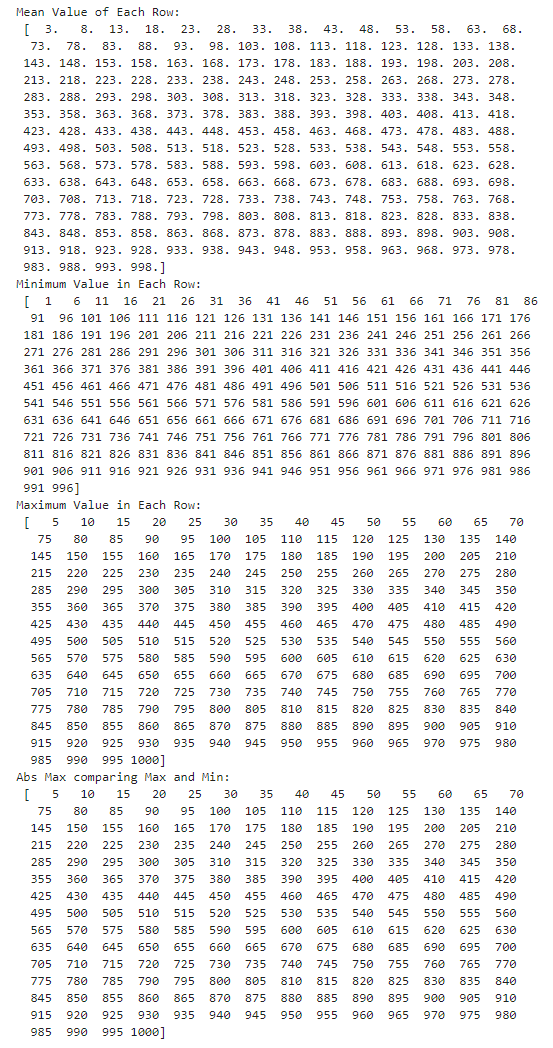
**Scope** is to create a python program that can generate 1000 random numbers and find the mean of every 5 numbers.   
Additionally, derive the minimum and maximum number in each set of 5 numbers.

**Solution**:   
Using NumPy function ‘arange’, generate 1000 random numbers starting from 1 with incremental 1.

Using the generated number range, split it into 5 columns and 200 rows using NumPy function ‘reshape’.

Using the NumPy function ‘mean’, ‘min’, ‘max’ and ‘maximum’, derive the mean value, minimum, maximum and Abs maximum value in each row.

The result is as below,



Data science is a branch of computer science where we study how to store, use and analyse data for deriving information from it. In order to do that, it is very vital to understand the importance of numbers in each and every dataset. To do that, we need to perform many arithmetic operations and the above-mentioned operations are very vital in that.

# Connect to MySQL DB, Create Table and Maintain data.

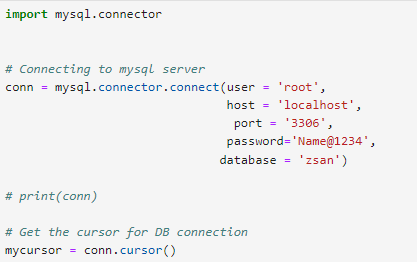
**Scope** of this task is to connect to a MySQL database from the python and create tables and maintain records from python.

**Solution**:  
By using the mysql.connector package, we can connect to the MySql database and using inbuilt functions like ‘cursor’ and ‘execute’, we can access the database tables in MySql, read and modify the data.

## Connect to MySQL Server

First step is to install the MySql server in the machine or have the necessary access details for remote database system.  
From jupyter notebook, import mysql.connector and using function ‘connect’, we can establish the connection.   
For a successful connection, it is important to provide the correct user credentials, host and port details.

When done, we will get reference to the MySql using which we can get the cursor for DB connection using function ‘cursor’.



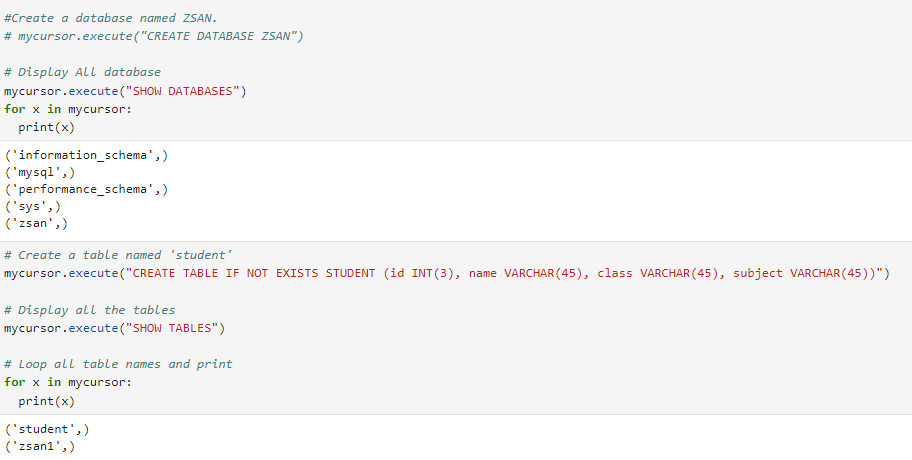
## Database and Tables

To make use of MySql, first step is to create a database under which all the tables can be stored.

Using the cursor variable and function ‘execute’, we can create or display the database using the below commands,

mycursor.execute(“CREATE DATABASE XXX”)  
mycursor.execute(“SHOW DATABASES”)

After creating database, we can create tables that are relevant for the project and work scenario.



## Insert and Select Records

Once the database and tables are created, next step is to maintain the data and the common activities are creating data and selecting data.

To do that, we can use the cursor execute function by passing the right database query.

Result shown below,



SQL plays a central role in data science and its ability to integrate effortlessly with modern tools, handle a variety of data operations, and support real-time processing, AI/ML, and data security makes it indispensable.

Data is a valuable business asset that continues to grow in importance. SQL serves as the language of data, integral for unlocking its power for data scientists, making it the most versatile tool for connecting data with data science. SQL’s simplicity and versatility, ability to integrate with modern tools, and its essential roles in artificial intelligence, machine learning make SQL indispensable to data science. (Shambavi Sivaramakrishnan)

# Exploratory Data Analysis (EDA) on Iris Dataset

**Scope** of this task is to conduct a detailed data analysis on the Iris dataset.

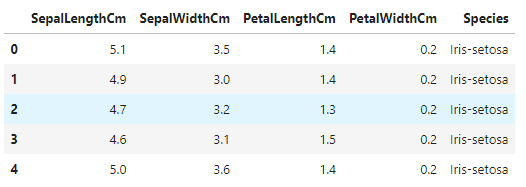
**Solution**:

For this study, Iris dataset from scikit-learn.org has been downloaded in csv file.

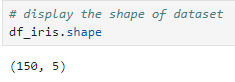
Iris dataset consists of 3 different types of irises’ (Setosa, Versicolour, and Virginica) petal and sepal length, stored in a 150x4 rows and coloumns.

The rows being the samples and the columns being: Sepal Length, Sepal Width, Petal Length and Petal Width.

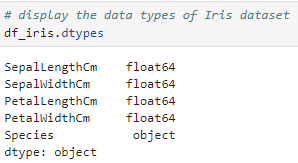
Import the dataset using panda functioin (read\_csv), followed by displaying the top 5 rows using function ‘head’.



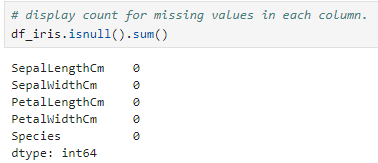
The shape of the dataset can be displayed using function ‘shape’. This helps in identifying the number of rows and columns.



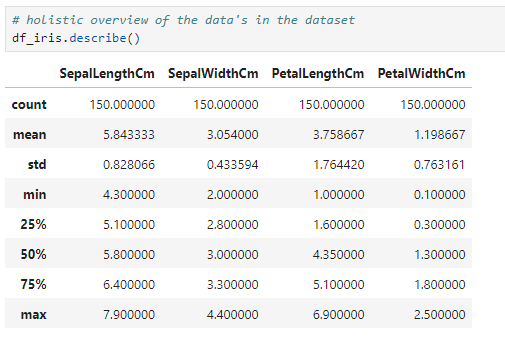
The datatypes of the columns can be identified using ‘dtypes’ of Panda library. This tells what type of data is stored in each column.



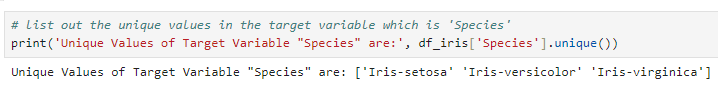
In order to check for any missing values in each column, we can use the function ‘isnull’ and to get the count of missing value, we can use function ‘sum’ of pandas library.



Complete overview of the dataset can be observed using function ‘describe’. This will return the total count, mean, std, min, 25%, 50%, 75%, max values for each column.



Unique values of specify column can be found using ‘unique’ function of pandas. In a very large dataset, this helps in identifying unique values in each column which is very important to list the possible values.

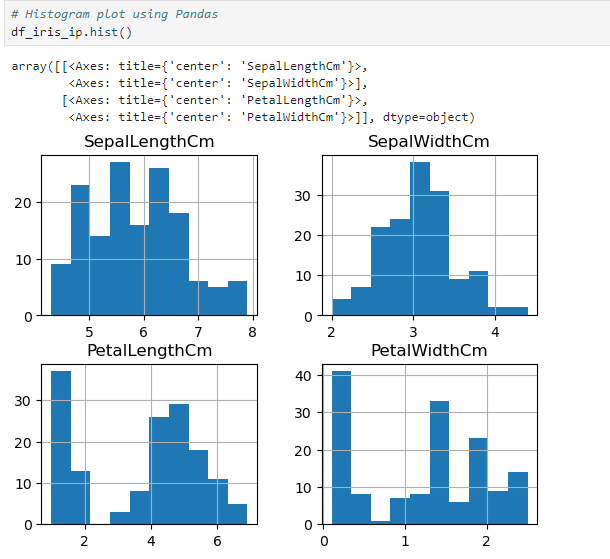


**Histogram**:

Histogram graphs are very helpful in identifying the frequency distributions and number of observations within each given interval. For this purpose, we shall use the ‘hist’ function from pandas. Some of the usual optional parameters of the ‘hist’ function are,   
column – Used to filter the columns in dataset for which the graph is returned

Xlabelsize, Ylabelsize – Both helps in changing the X and Y axis label size.

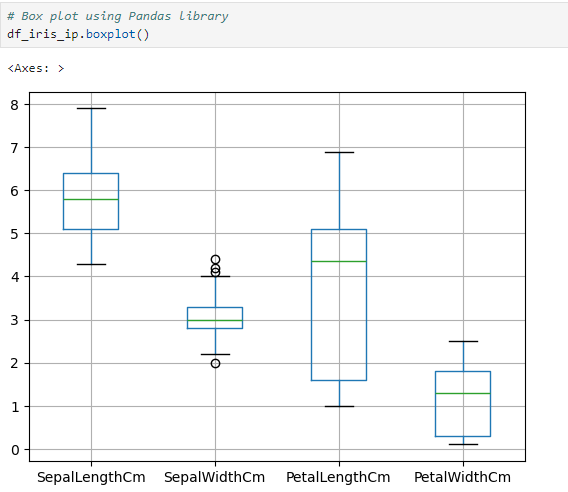
Result is as below,



X axis contains the value of each feature and Y axis contains the occurrence.   
This graph clearly shows that petal width ranges from less than 1 to under 3. Also, sepal width has high occurrences at approx. 3.  
We can see that sepal length value is more than 4 and goes above 7.

**Boxplot**

The boxplot helps in summarizing the distribution of a numeric variable for one or several group. It allows to quickly get the median, quartiles and outliers.  
Boxplots can be made with matplotlib, seaborn library as well but here in this case, we will use ‘boxplot’ function from Pandas library.



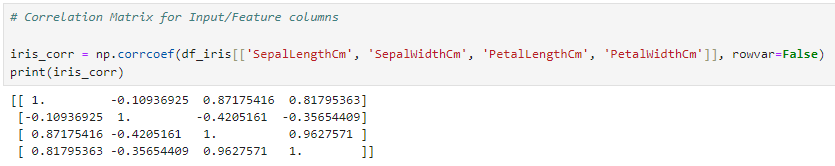
Result from Boxplot matches with the findings from histogram. Sepal width has high occurrences at value close to 3.

**Correlation Matrix**

Correlation matrix helps in identifying the relationship between 2 features. Correlation value of 0 indicates no relationship between 2 features. Correlation value of 1 indicates both are strongly proportionate relationship. This means that value for both features increase or decreases at a constant ratio.

Correlation value of -1 indicates that both are disproportionate to each other.

This helps in model building and to predict many scenarios.#



This result tells that there is a strong positive correlation between Petal length and Petal Width. And a strong negative correlation between sepal width and petal length.

# References

(Ashraf)

Ashraf, Abdallah. “Correlation in Machine Learning — All You Need to Know.” *Medium*, 22 Sept. 2023, medium.com/@abdallahashraf90x/all-you-need-to-know-about-correlation-for-machine-learning-e249fec292e9.

Available at:

<https://medium.com/@abdallahashraf90x/all-you-need-to-know-about-correlation-for-machine-learning-e249fec292e9>

(Shambavi Sivaramakrishnan)

Shambavi Sivaramakrishnan. “SQL in Data Science: A Timeless Tool in a Modern World.” *RTInsights*, 29 July 2024, [www.rtinsights.com/the-relevance-of-sql-in-data-science-a-timeless-tool-in-a-modern-world/](http://www.rtinsights.com/the-relevance-of-sql-in-data-science-a-timeless-tool-in-a-modern-world/).

Available at:  
<https://www.rtinsights.com/the-relevance-of-sql-in-data-science-a-timeless-tool-in-a-modern-world/>

# GitHub Link

<https://github.com/santhosh-sba24100/CA1-Python_Programming>