Course of Study: MSc Computer Science On-Campus  
Type of Thesis: Research & Practical  
Date: 9th June 2024  
Author Name: Abdul Ghani Chishti  
Author Email: abdul-ghani.chishti@iu-study.org  
Matriculation Number: 4242853  
Tutor Name: Cosmina Croitoru & Tai Le Quy

Data Analysis using python libraries in this era vs without libraries

Course Name: Programming With Python DLMDSPWP01

Table of Contents

[**1. INTRODUCTION** 2](#_Toc168848270)

[**2.DATA ANALYSIS USING PYTHON WITHOUT LIBRARIES** 3](#_Toc168848271)

[**2.1. Tabular data.** 3](#_Toc168848272)

[**2.2. Manual Calculations** 3](#_Toc168848273)

[**2.3. Graphical Representation.** 4](#_Toc168848274)

[**2.4. Iterative Processes.** 5](#_Toc168848275)

[**2.5. Sorting Algorithms.** 5](#_Toc168848276)

[**2.6. Data Structures** 5](#_Toc168848277)

[**3.PYTHON LIBRARIES FOR DATA ANALYSIS** 5](#_Toc168848278)

[**3.1. OVERVIEW OF PYTHON LIBRARIES** 5](#_Toc168848279)

[**3.1.1. Numpy:** 5](#_Toc168848280)

[**3.1.2. Pandas:** 6](#_Toc168848281)

[**3.1.3. SciPy:** 6](#_Toc168848282)

[**3.1.4. Scikit-learn:** 6](#_Toc168848283)

[**3.1.5. Matplotlib:** 6](#_Toc168848284)

[**3.1.6. Seaborn:** 6](#_Toc168848285)

[**3.2.DATA ANALYSIS USING PYTHON LIBRARIES** 7](#_Toc168848286)

[**4. COMPARISON BETWEEN DATA ANALYSIS USING PYTHON LIBRARIES AND WITHOUT LIBRARIES** 10](#_Toc168848287)

[**4.1. Detail Comparison** 11](#_Toc168848288)

[**4.2. Future of Data Analysis using Libraries** 12](#_Toc168848289)

[**5.CONCLUSION** 13](#_Toc168848290)

[**6. Relation between Research Work & Practical** 13](#_Toc168848291)

[**7.REFERNECES**: 14](#_Toc168848292)

[**8. APPENDIX** 15](#_Toc168848293)

[**8.1 About Code** 15](#_Toc168848294)

[**8.2 Code** 15](#_Toc168848295)

[**8.3 RESULT** 20](#_Toc168848296)

[**8.4 VISUALIZATION** 20](#_Toc168848297)

# **­1. INTRODUCTION­­­­**

In the era of data and where data is the key for everything, the importance of efficient data analysis cannot be overstated. Data analysis is easier now a days because of various libraries available in python programming language for doing all sorts of analysis as compared to the era when there were no such libraries available. Data Analysis using python before the libraries were introduced was such a hectic task as the size of the data increased the amount of effort it took to analyze, clean and perform statistical methods takes a lot of time.

The primary objective of the research is to evaluate how Python libraries changes the data analysis processes, improve the efficiency and its accuracy. The traditional methods used earlier before the python libraries evaluation and its process will be discussed. The different techniques and analysis procedures that researcher take to complete the data analysis and the research will also focus on the methodologies that were used using different python libraries such as (Numpy, Pandas, Matplotlib) and prior so that there will be clear distinction between the efficiency of the output of both libraries and without libraries.

The discussion will not delve into the structure of the libraries or types of algorithm used in the libraries but instead the focus will remain on the practical implementation and the advancement the libraries offers. Additionally, the old methods before libraries will also be discussed to give the gest of both of them equally. Furthermore, Finally the comparative analysis will highlight how the traditional methods and python libraries enhance the data analysis process to get the clear understanding of how python programming language libraries revolutionize the data analysis.

# **2.DATA ANALYSIS USING PYTHON WITHOUT LIBRARIES**

Data analysis when modern libraries are not introduced were difficult and requires more time than it can be done now in seconds. When there were no libraries for data analysis each and every calculation should be performed manually in code. For each statistical or scientific computation a manual code is written, to start the data analysis all the numerical, statistical, scientific methods are created that will be used on the data. On the other hand, data is manually handled using basic fileing procedures that will read or write the file (if using csv or xls). This requires another piece of code that will handle the data that will accurately read the file and than on that data the analysis is performed. All this process has to be done in order to perform a simple data analysis but if the data is too large than it will takes a lot of time.

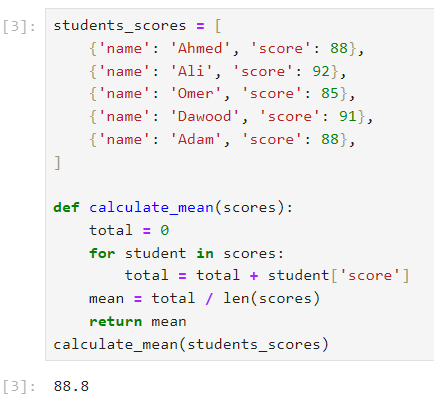
However, Traditional technique for data analysis was used when there were no such advance libraries available in python programming language. Every data scientist use this technique to start the data analysis. It is as follows:

## **2.1. Tabular data.**

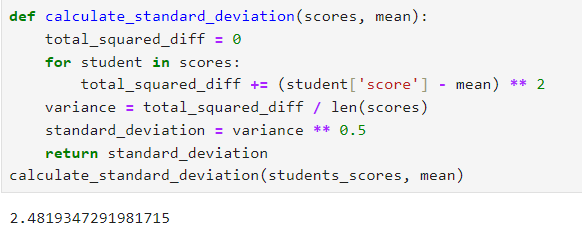
The Data can initially be in raw form or unstructured. For data analysis, the data is converted into tabular format so that it will be easily readable and can have clear understanding of how the data will be in structured format. Some data can already be in a structured format that will minimize the time of converting the data into structured form. The tabular data is often used for small data sets for clear data representation. Some data can be analyzed without tabular format as well.

## **2.2. Manual Calculations**

To perform the necessary calculation in python when there were no libraries, all functions has to be made manually using arithmetic and logic expressions. The basic statistical functions also be made manually consuming extra time and making the process of data analysis slower. The basic statistical methods used in early stages of data analysis are mean, median, mode, standard deviation etc. These basic functions were made manually to perform statistical processes on data.  
  
Using arithmetic and logical operators of python programming mathematical or statistical functions can be made. Here is the code snippet of some functions in python without using any advance libraries:

Fig.1 Calculating Mean of the data without using libraries

Source: own representation

Fig. 2 Calculating Standard Deviation

Source: Own representation

## **2.3. Graphical Representation.**

Graphical representation of the data helps to understand the data more easily and effectively but when there was no advance python libraries for visualization the sketches were drawn manually or by using drawing tools in python.

## **2.4. Iterative Processes.**

In conventional data analysis, iteration is a basic method. A fundamental technique is looping through data points to carry out calculations, update counts, or verify conditions. For tasks like figuring out modes, summing frequencies, and computing means, this iterative method is crucial for data analysis so that one procedure can be repeated if needed.

## **2.5. Sorting Algorithms.**

For many data analysis tasks, such determining medians or generating frequency distributions, sorting is an essential first step. Sorting algorithms such as bubble sort, insertion sort, or quicksort can be implemented with the help of simple Python structures. This enables ordered data processing, which is essential for many statistical computations.

## **2.6. Data Structures**

Traditional data analysis requires the use of fundamental data structures like lists, dictionaries, and tuples. Lists are used for managing immutable data, dictionaries for keeping frequency counts, and lists for storing datasets. Gaining proficiency with these structures makes it more effective and efficient data analysis and manipulation.

# **3.PYTHON LIBRARIES FOR DATA ANALYSIS**

Data analysis is become more crucial in many industries that works on large datasets or the industries whose work are data driven. Python program language with its extensive libraries has make the data analysis easier and efficient. From at least two decade, python language has gone in advance scientific computing *(McKinney, W. (2012), p.2)* and simple to handle that makes it the good choice for the data analysis. Python libraries provides various functionalities such as data manipulation, statistical analysis and visualization. Python libraries that are widely used are Numpy, Pandas, Scipy and Matplotlib. These libraries helps in data analysis of the data that are too large to be handled manually. Some library also provides graphical representations for the data that helps the data scientists and researchers for better understanding.

## **3.1. OVERVIEW OF PYTHON LIBRARIES**

Python has vast variety of libraries and data analysis there are all sort of libraries and even alternative libraries that can be used for same purpose. Here are some key libraries that are widely used in data analysis.

### **3.1.1. Numpy:**

Numpy (Numerical Python) is a package used for extensive numerical computations. To perform any sort of computation in python numpy library is used as It can provide great efficiency and gives faster output. It can also do numerical computing on multi-dimensional data like arrays and objects. [(Saabith, A. S., Vinothraj, T 2020, p24).](#_6.REFERNECES:)

### **3.1.2. Pandas:**

Pandas (Python Data Analysis) is most popular library it widely used for the purpose of data manipulation alongside with numpy, and Scipy. Pandas is used for large volume structured data whose size goes around in gigabytes. [(Stepanek, H. (2020.p1)](#_REFERNECES:) It also provide the feature of handling the missing data in the dataset if any. For real world data analysis, pandas is used as it has high performance and can be used easily without any complications.

### **3.1.3. SciPy:**

SciPy (Scientific Python) it is a collection of packages used for scientific computations and calculations. SciPy also contains statistical methods that helps to calculate the statistics of the data. SciPy is very helpful for real world data analysis. Some of its basic functions are stats, optimize, integrate etc [(McKinney, W. (2012), p.6).](#_REFERNECES:)

### **3.1.4. Scikit-learn:**

Scikit-learn is a machine learning package in python programming language. Thanks to its ease of use structure and performance that helps data scientist and researcher to utilize it for data driven work and analysis. Scikit-learn also provide graph visualization for easy analysis. [(Bonald, T., De Lara, N., Lutz, Q., & Charpentier, B. 2020, p1).](#_REFERNECES:)

### **3.1.5. Matplotlib:**

Matplotlib is widely used for generating plots, graphs and two or three dimension visuals for the data. Matplotlib provides all sorts of output that can be used by data scientist. It used along side with pandas together it will create a better lineup for data analysis and better output. [(Saabith, A. S., Vinothraj, T 2020,p24-25 )](#_REFERNECES:)

### **3.1.6. Seaborn:**

Seaborn is a visualization library built by using matplotlib. It provides basic visualizations of the data that are frequently used. Seaborn is mostly used with Pandas library data frames. Plots that are made by seaborn are often similar with matplotlib. [(Sial, A. H., Rashdi, S. Y. S., & Khan, A. H. 2021, p2)](#_REFERNECES:)

These are some most popular python libraries used for data analysis now a days and these libraries also providing assistance for machine learning. By using these libraries a data scientist can easily perform its numerical. scientific and statistical analysis is short period of time with less error chances and fast outcomes. Even with the graphical and plot representations of the data alongside the calculations.

## **3.2.DATA ANALYSIS USING PYTHON LIBRARIES**

Data analysis is crucial for the data scientists and researchers. Using python libraries for data analysis makes it easier, faster and more efficient. By these libraries data analysis is simpler from manually calculating and analyze the data. Data scientist often wants to perform various calculations that takes a lot more time and by using these libraries the results are produced in seconds, making it time efficient as well. Earlier there were no big data available that requires a lot of processing before the analysis but in this era every thing rely on data and the size of data is increasing for each sector exponentially and to analyze huge data the old tactics will not work accurately. Using python for analysis will not only boost the outcome efficiency but also minimize the human work load to calculate and evaluate all the data by themselves even using a programming language it takes time if any library is not used along with the analysis.

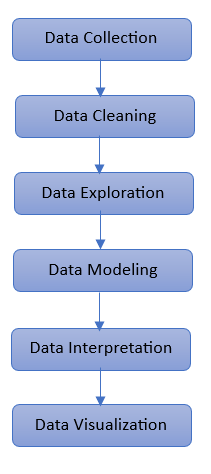
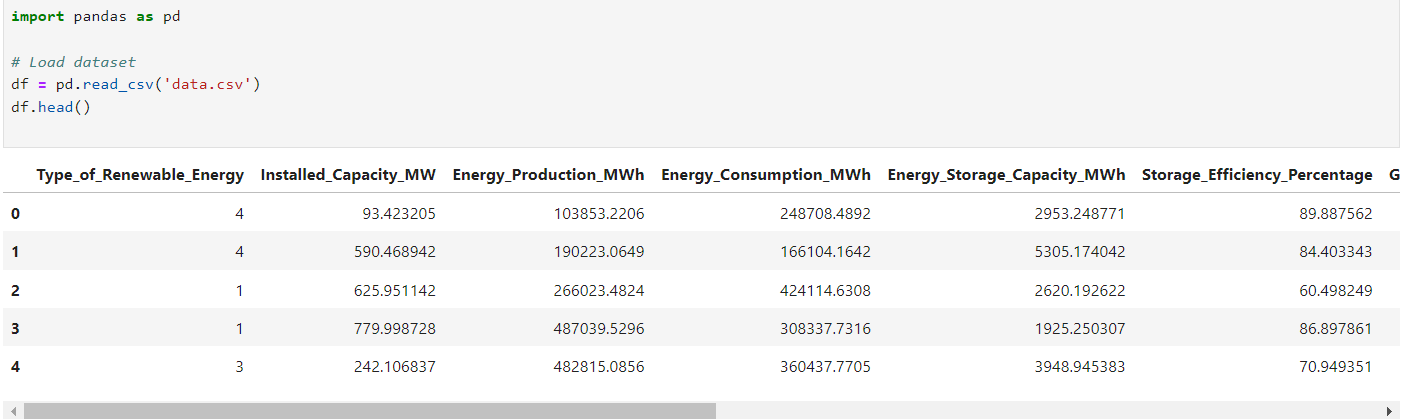
The data analysis using Python have some series of steps to follow like a workflow that has to be done in order because skipping any step can lead to inaccurate results. The basic data analysis workflow using python according to (www.polymersearch.com) is shown in figure 1

Fig.3 Data Analysis workflow

Source: Own representation

The above figure shows the steps that is used in data analysis. Starting from the Data Collection which basically can be databases, csv (comma separated file), excel files etc. These form of data files are used. In python these files are loaded using pandas library that are than further processed. Processed means handling of empty or null data. Here is the example of data collection using pandas and creating a data frame

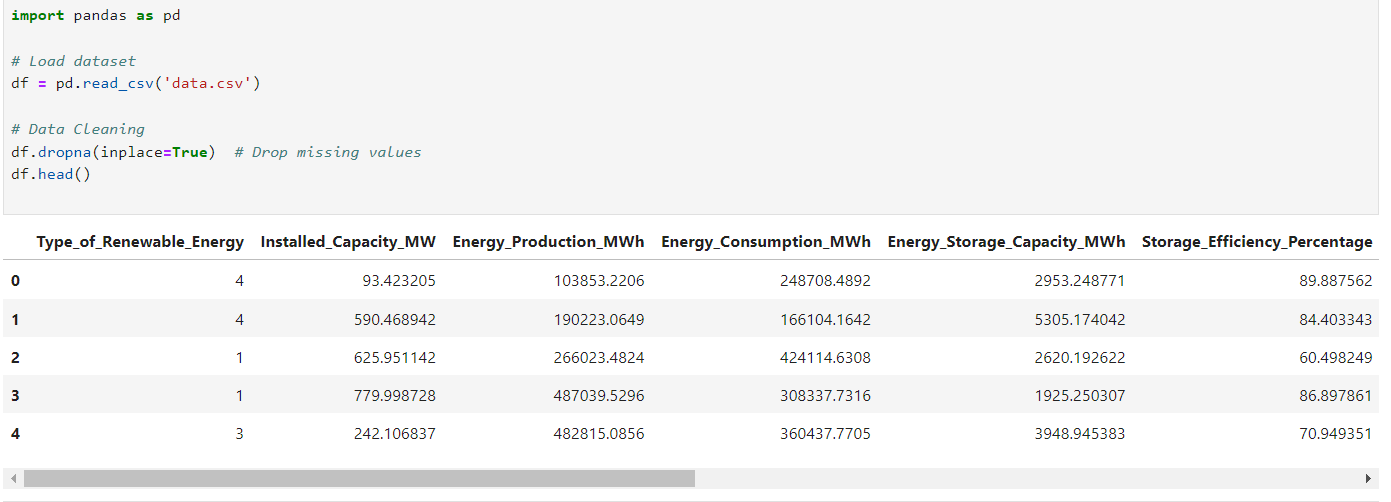
Fig.4 Data Collection using Pandas



Source: Own representation

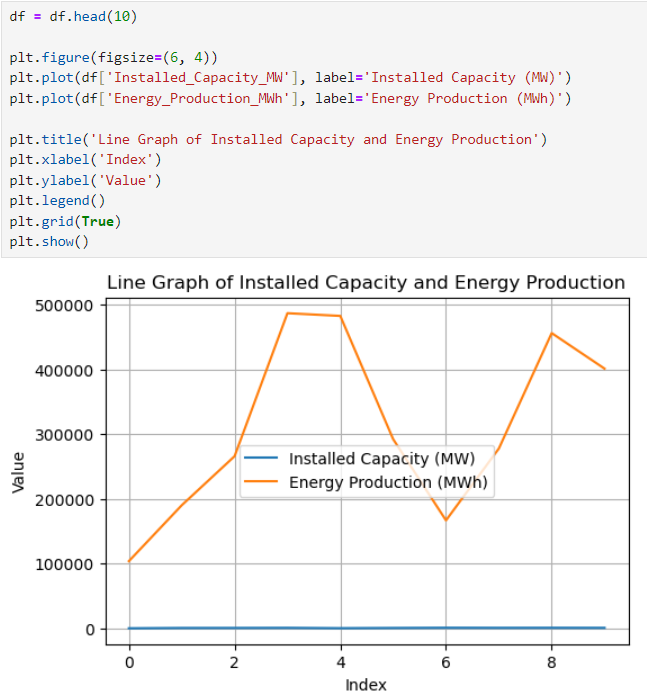
After data collection, if there is any null values it can create ambiguity in the output so it has to be handled. For null or not available records in the data it can be dropped from the data for precautionary measure. By using Pandas library it makes it easier to handle null values just by using single line of code it is done. It is considered to be cleaning of the data. Here is the code snippet for the dropping of null values from the data.

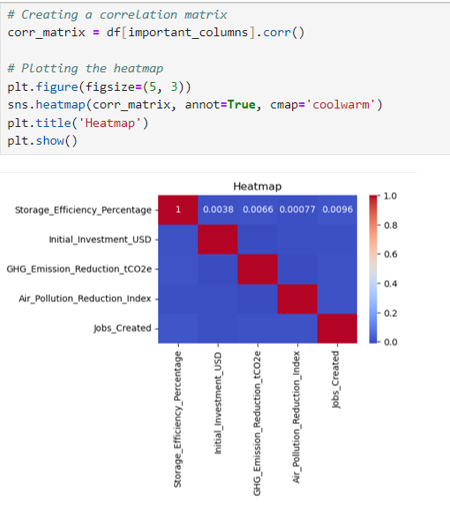
Fig.5 Dropping Null values from data



Source: Own representation

Now after data cleaning, data exploration, modeling, interpretation can be done according to the data scientist. After that data analysis final step is for data visualization can be done. There are many types of graph that are used by data scientist for data analysis some are heatmap, line graph, histogram etc. These graphs are made by using matplotlib library. Here is the different types of visualizations using matplotlib in python.

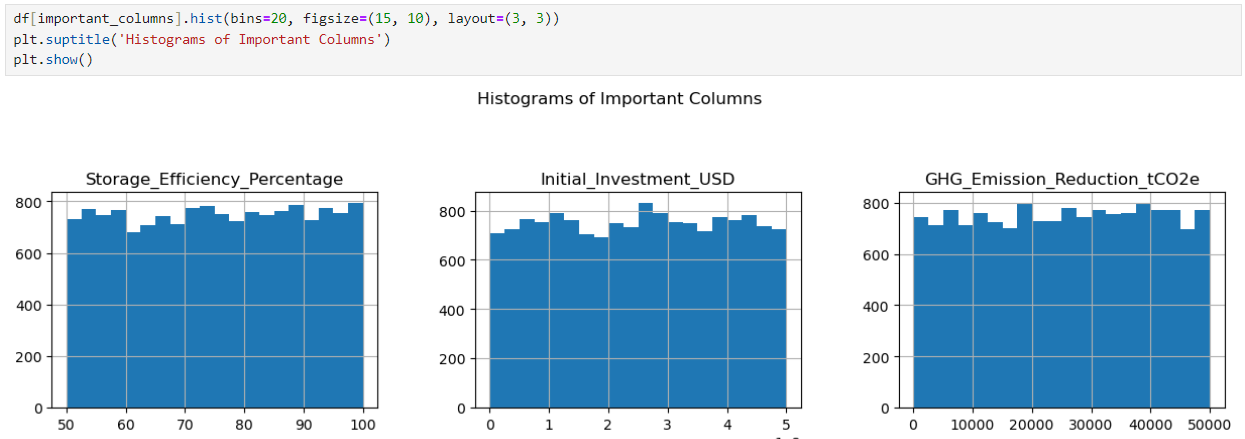
Fig.6 heatmap Fig.7 Line graph



Source: Own representation

Source: Own representation

Fig.8 Histogram



Source: Own representation

These are some of the data visualizations for the data analysis using matplotlib python library. There are many more graph and charts available to use. Making graphs and chart become very simple just by using matplotlib and it is saving the time for creating the graphs manually like traditionally used before these libraries were launched. [(International Journal of Innovative Technology and Exploring Engineering (IJITEE),2019, p.3)](#_REFERNECES:)

# **4. COMPARISON BETWEEN DATA ANALYSIS USING PYTHON LIBRARIES AND WITHOUT LIBRARIES**

Data analysis using python without libraries is not being in practice since there are many advance python libraries that are available that facilitates the data scientist to efficiently work on data with all sort of computation and visualizations. While before these libraries the data manipulation, cleaning and visualization takes longer time and have chances of some errors due to manual computation is performed on the data. Using Python libraries can also minimize the human efforts and the time it takes to each and every step manually and make separate custom functions for it will also be replaced. The table below shows the side by side qualitative comparison for both data analysis using python with or without libraries.

Table.1 Data analysis with python vs without python

|  |  |  |
| --- | --- | --- |
| **Aspect** | **With Python Libraries** | **Without Python Libraries** |
| Efficiency and Speed | Fast Execution | Slow Execution |
| Data Loading | Efficient | Slow |
| Data Manipulation | Fast Execution | Slow Execution |
| Statistical Analysis | Fast, reliable | Slow, higher computation |
| Visualization | Efficient, Easy | Manual drawing |
| Ease of use | Easy to write and understand | Require extensive code |
| Dependencies | Require installation of external library | No dependencies |
| Development time | Rapid development | Time consuming |
| Educational Value | Good for learning | Excellent for understanding concepts |

Source: Own representation

## **4.1. Detail Comparison**

According to [Li Jun, & Li Ling. (2010)](#_REFERNECES:) python has increased its processing speed in every releases libraries like pandas, numpy and scipy are used on large size datasets. Loading data using pandas is efficient as it is designed to handle large file swiftly. While with old traditional methods are significantly slower for example: manual reading a csv file and parsing it line by line will take a lot much longer because reading files involve iterations over each line and splitting it makes it slower and prone to error. So using a library that will be a better option while doing data analysis using python language.

Data manipulation is one of the fundamentals of data analysis and by using pandas for high level functions like grouping, merge, aggregate become easier and faster as the data manipulation without library require custom functions and explicit loops making it complex and slower. According to [Bröker, O., Chinellato, O., & Geus, R. (2005, p.972](#_REFERNECES:)) the complexity of the code and calculations makes it slower to execute and hence using python libraries for data manipulation is better option rather than making custom functions manually. Other than that, statistical analysis using python libraries are also vary efficient. The commonly used library for statistical analysis is scipy. [(Nunez-Iglesias, Van der Walt, & Dashnow, 2017, p.223)](#_REFERNECES:). Its output is faster and reliable while if statistical analysis is done manually the chances of errors become higher. According to [Cai, X., Langtangen, H. P., & Moe, H. (2005, p.55)](#_REFERNECES:) the numerical computing in python without library like numpy is making the execution a lot more slower as the complexity and length of the numerical computation is, so library is preferable.

Data Visualization is key for better analysis as it provide the summarized outcome from the statistical analysis. The most widely used data visualization library in python language is matplotlib and it provide all sorts of graphs, charts. Its use is straightforward and easy to use. [(ABDİMANAPOVA, G., ZHAİDAKBAYEVA, & ALDESHOV, S.2023 p1)](#_REFERNECES:). The line of code will be minimal and it can process large data aswell. While manually drawing graphs and charts require excessive time and precision to be accurate as any minor mistake can change the output of the analysis as the data analyst will go on with the error output. Furth more, the libraries that are used in data analysis cannot be used without installing it into the development environment. Each and every library has to be installed before it can be used in the program. Without any installed library nothing will work. It might cause dependency issues in some environments. On the other hand, there is no dependencies are required if there is no use of external libraries which will not cause any dependencies issues in the data analysis work.

Development time for data analysis using python is much faster than that of the without one because all the functions that are used are not required to be written manually while without library work will takes time for creating each functionally manually. But for concepts it is preferable the data analyst should practice manual before using libraries.

## **4.2. Future of Data Analysis using Libraries**

Data Analysis is getting advanced day by day and in some years, it will be more advanced. This modern era using data for each and everything and size of data is exceeding to handle it the upgradation is required for everything. Now for its future the data analysis will be also taking over by the artificial intelligence to make the data analysis more efficient and accurate than ever before. In near future it may be possible that python library will provide full three dimensional visualization and real time data analysis with the help of machine learning libraries in python.

The vast advancement of technology will also impact how the data analysis is performed and how can it be done without any human interactions to avoid human error and to get 100% accurate results. This will take a lot of time happen as the artificial intelligence is still in its development stage to take over whole data analysis and statistical analysis by itself. But some form of AI managed data analysis is still available which is again based on data. The future scope of the data analysis is illustrated in the figure below

Fig.9: Data analysis future scope

Data Loading

Data Cleaning using advance libraries

Data visualization using AI python model

# **5.CONCLUSION**

A good understanding of statistical ideas and basic programming concepts are necessary to perform data analysis in Python using conventional approaches without the use of libraries. Traditional methods provide important insights into the underlying mechanics of data processing and improve problem-solving abilities, even when modern libraries offer practical and effective tools for data analysis. Gaining proficiency with these fundamental techniques can help one understand the intricacies of data analysis and the power of programming on a deeper level.

Python libraries are highly recommended for most practical data analysis jobs because of their efficiency, sophisticated functionality, and user-friendliness. Nevertheless, doing data analysis without libraries offers more insightful education and a more thorough comprehension of the underlying ideas, which can be useful in educational settings or in settings where external dependencies are strictly limited. By comparing the data analysis techniques using python libraries and without the libraries gives the conclusion that the using the libraries to perform the task of data analysis is beneficial as compared to the data analysis without using python library due to its time consumption and slow response time.

# **6. Relation between Research Work & Practical**

So, by considering the given datasets and the requirements of the practical I used multiple libraries to solve it which are explained above in detail and each library has its own responsibility like I used panda to read all the given dataset which is in CSV format, I used numpy to perform calculation based on mathematics. Also I used curve\_fit function from the scipy library to fetch out all the predicted ‘Y values’ by manipulating the given training dataset. On the other hand it is also possible to solve the practical without using the libraries but it is not a good approach because the number of **lines of code** would be too much and also the **complexity** of the functions will increase and the chances of **occurrence of bugs** will also increase. One more thing is that there is **no way in python to visualize** the data without using its library that’s why I used bokeh for this. This is how my research work is directly connected to the given data.

# **7.REFERNECES**:

[1]. McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".

[2]. Saabith, A. S., Vinothraj, T., & Fareez, M. (2020). Popular python libraries and their application domains. *International Journal of Advance Engineering and Research Development*, *7*(11).

[3]. Stepanek, H. (2020). *Thinking in Pandas*. Berkeley, CA, USA: Apress.

[4]. Bonald, T., De Lara, N., Lutz, Q., & Charpentier, B. (2020). Scikit-network: Graph analysis in python. *Journal of Machine Learning Research*, *21*(185), 1-6.

[5]. Sial, A. H., Rashdi, S. Y. S., & Khan, A. H. (2021). Comparative analysis of data visualization libraries Matplotlib and Seaborn in Python. *International Journal*, *10*(1), 45.

[6]. Nunez-Iglesias, J., Van der Walt, S., & Dashnow, H. (2017). *Elegant SciPy: The Art of Scientific Python*. " O'Reilly Media, Inc.".

[7]. Jun, L., & Ling, L. (2010, October). Comparative research on Python speed optimization strategies. In *2010 International Conference on Intelligent Computing and Integrated Systems* (pp. 57-59). IEEE.

[8]. Bröker, O., Chinellato, O., & Geus, R. (2005). Using Python for large scale linear algebra applications. *Future Generation Computer Systems*, *21*(6), 969-979.

[9]. ABDİMANAPOVA, G., ZHAİDAKBAYEVA, L., & ALDESHOV, S. (2023). GRAPHICS CAPABILITIES IN PYTHON. *Avrasya Sosyal ve Ekonomi Araştırmaları Dergisi*, *10*(1), 1-14.

[10]. Cai, X., Langtangen, H. P., & Moe, H. (2005). On the performance of the Python programming language for serial and parallel scientific computations. *Scientific Programming*, *13*(1), 31-56.

# **8. APPENDIX**

## **8.1 About Code**

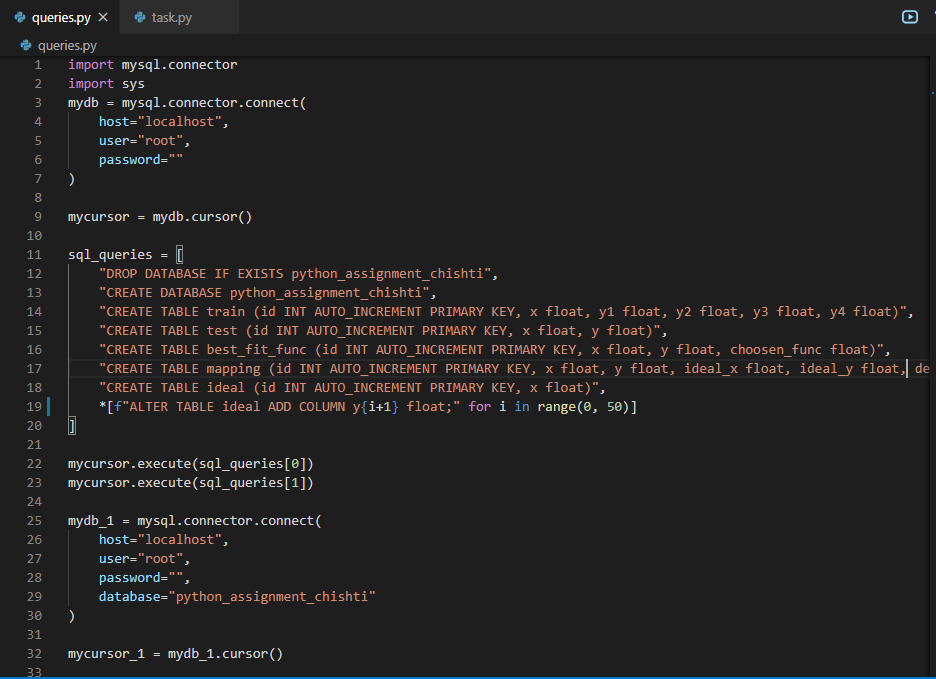
Github Link: <https://github.com/abdul-ghani-chishti/programming_with_python_assignment.git>

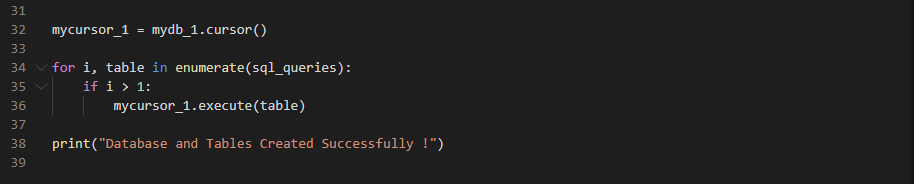
Note: The practical work have two python files the first one named queries.py and task.py

1. Queries.py -> is responsible to create database if not exist and also to create tables.
2. Task.py -> it is a main file which have all the operations.
3. Used <https://docs.bokeh.org/en/latest/> for visualization.
4. Used <https://www.w3schools.com/python/> for code as a reference.
5. Used <https://pythonnumericalmethods.studentorg.berkeley.edu/notebooks/chapter16.04-Least-Squares-Regression-in-Python.html> for least square method.

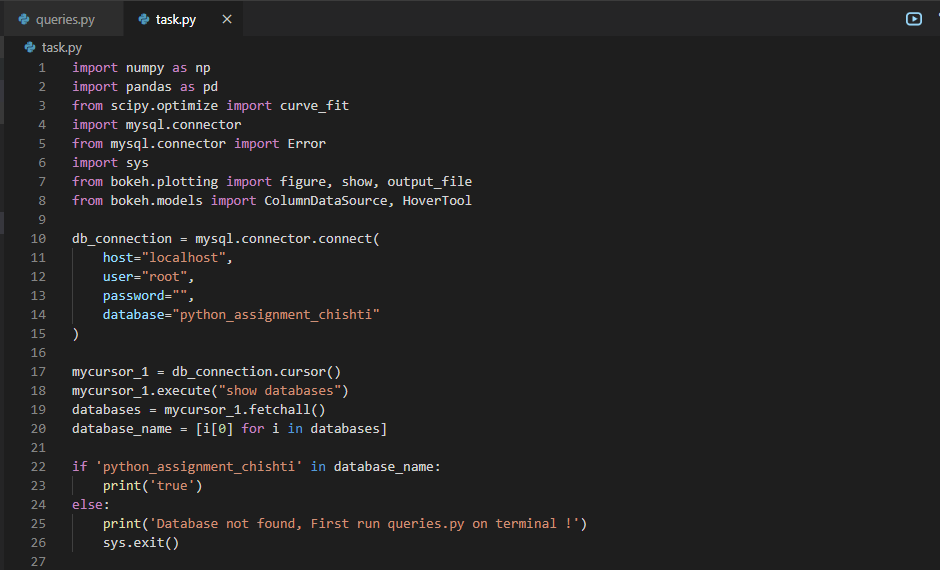
## **8.2 Code**

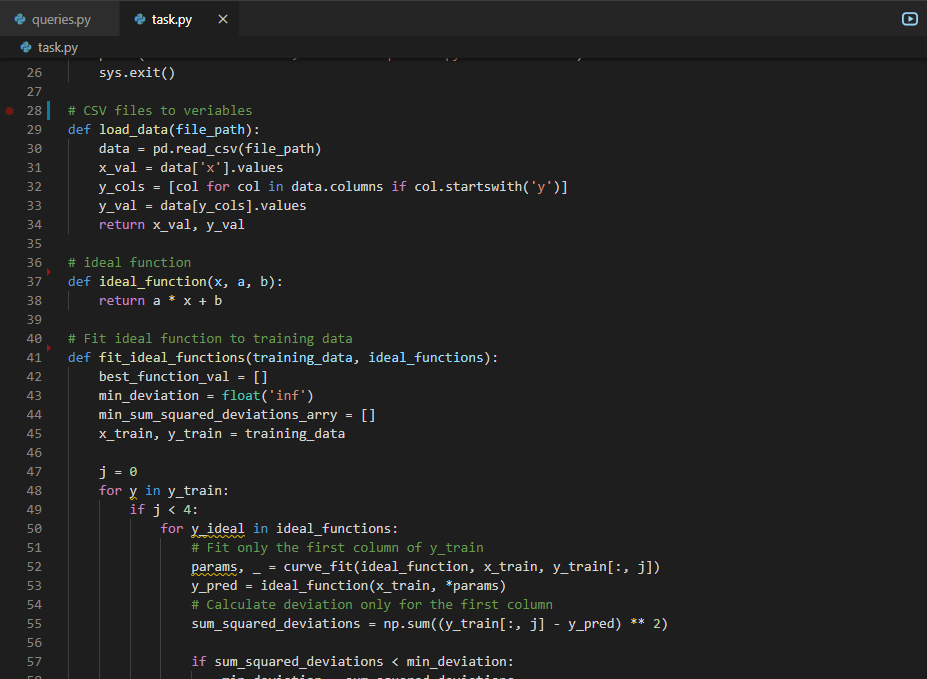
**Queries.py**

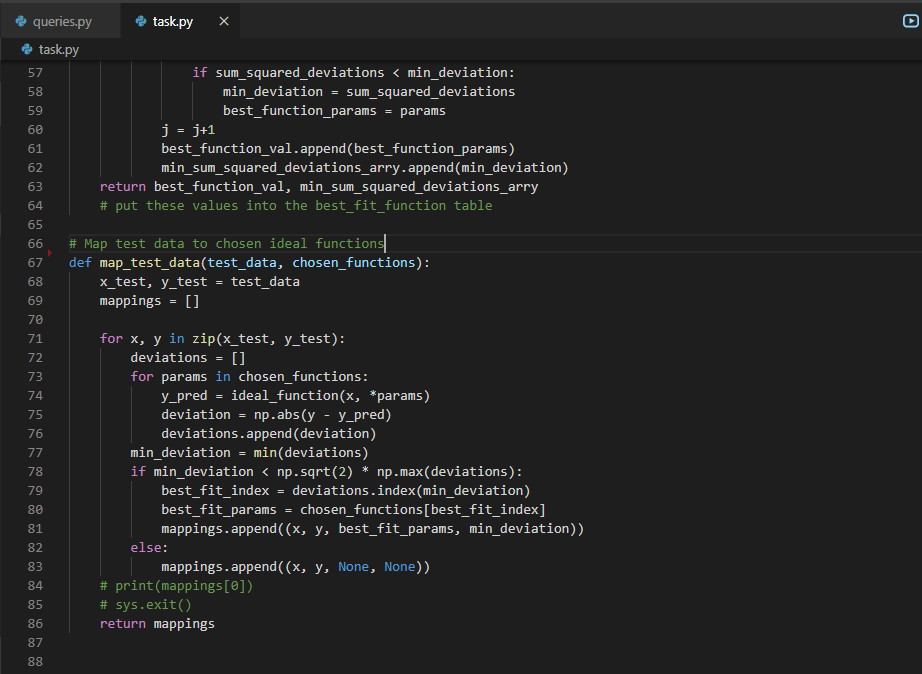


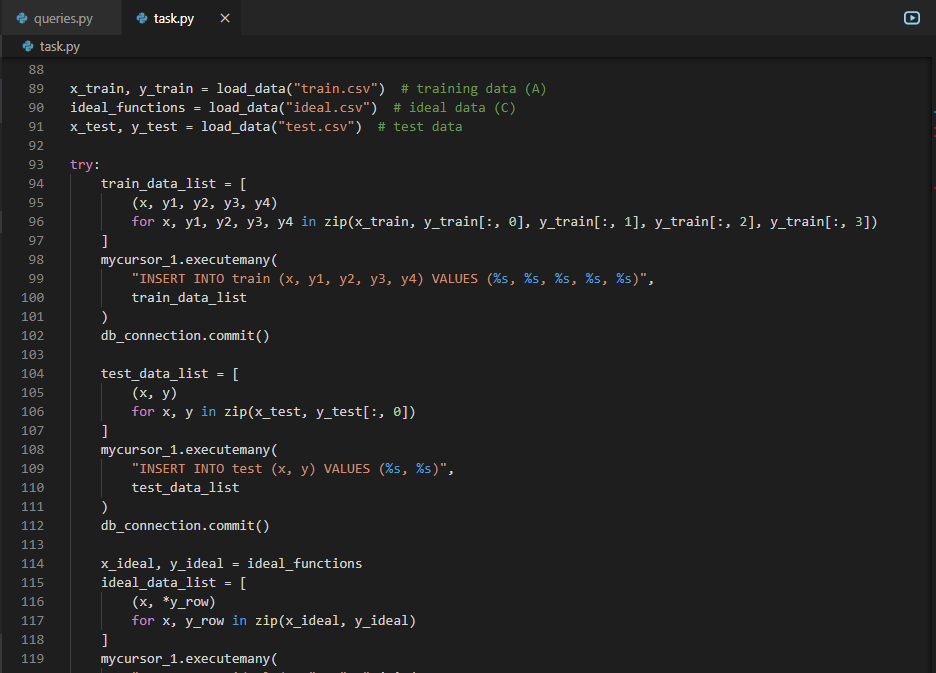


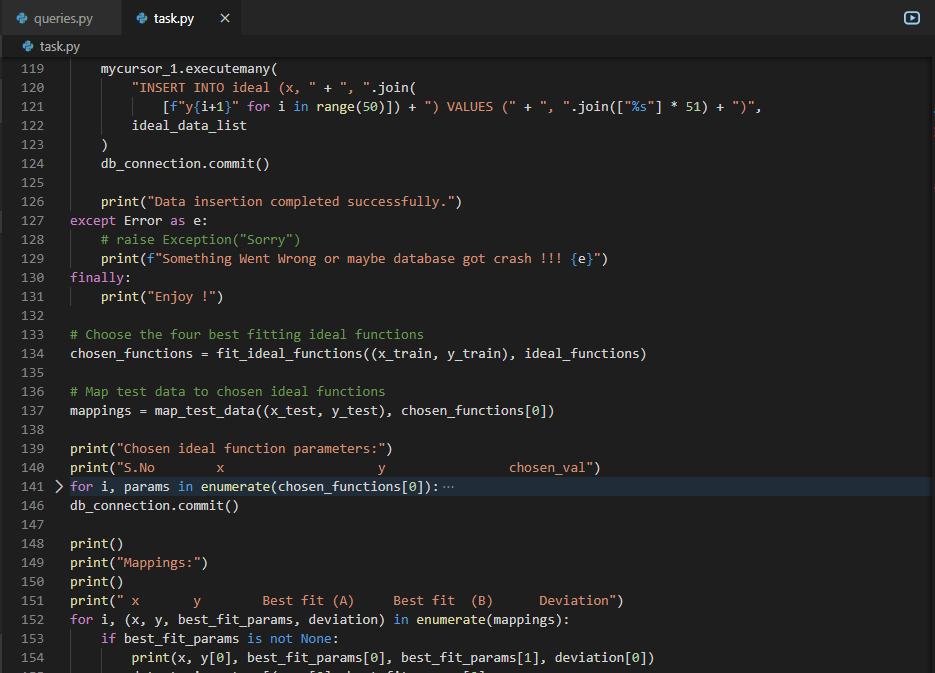
**Task.py**

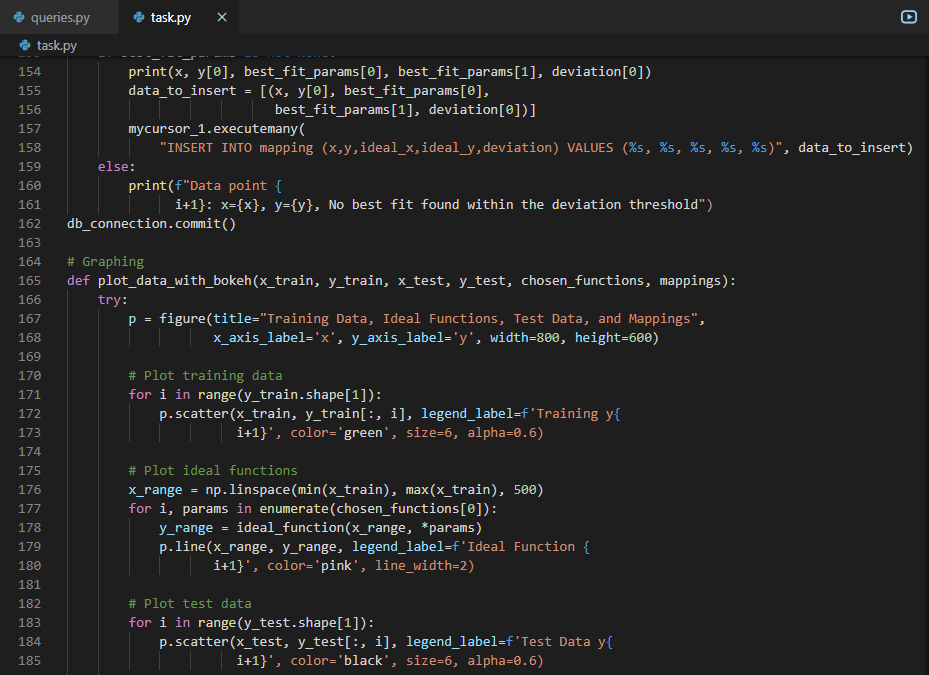


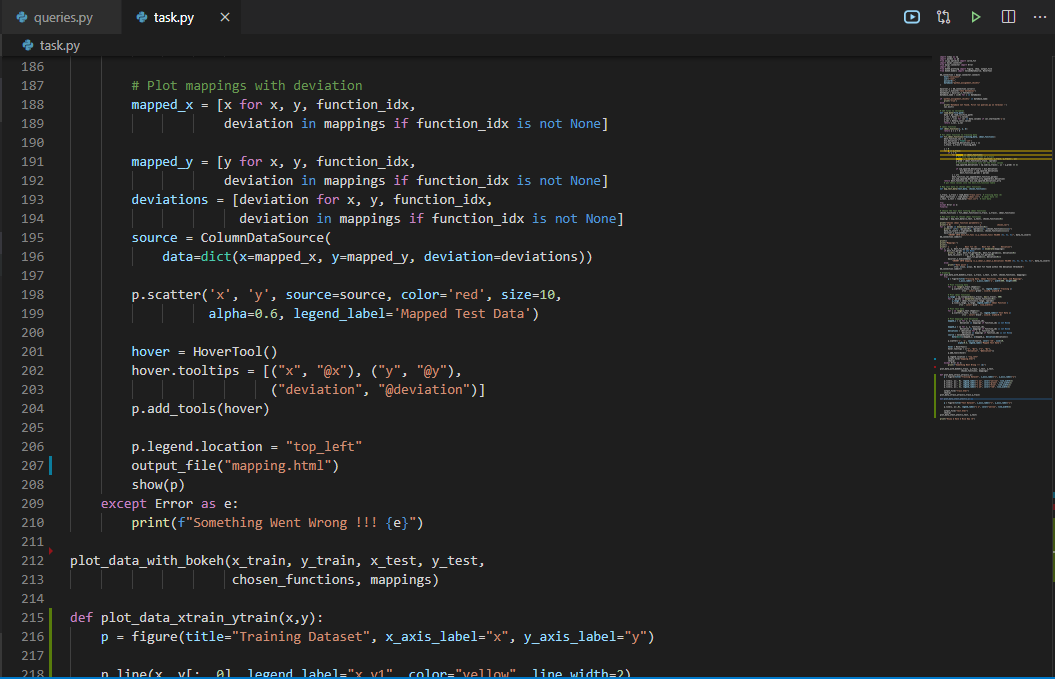


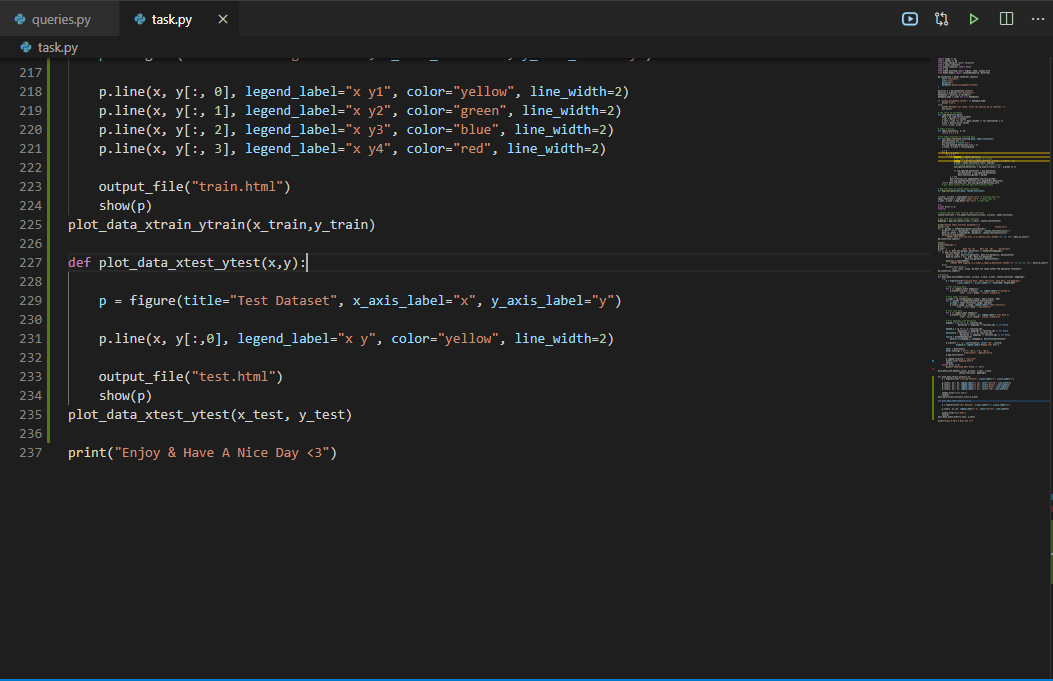




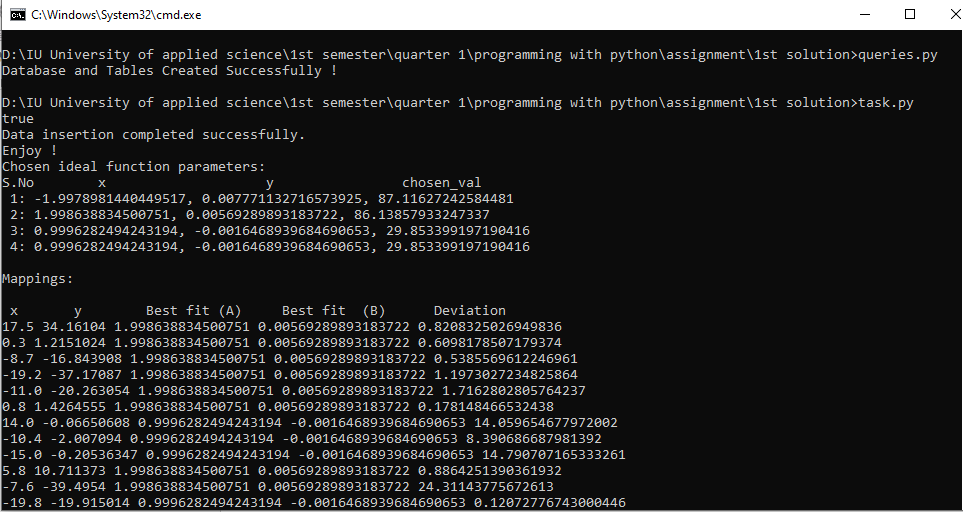








## **8.3 RESULT**



## **8.4 VISUALIZATION**

