



EMPIRICAL STUDY ON THE IMPORTANCE OF PYTHON LANGUAGE IN DATA SCIENCE

Jemima Hossain

**12th CBSE, APDS (Advance Programme in Data Science)
West Bengal, India.**

ABSTRACT

Data Science is a combination of interdisciplinary field that combines, programming, artificial intelligence, mathematics, scientific methods, statistical techniques, and data analysis to extract meaningful information from data and actionable knowledge from structured and also from unstructured data. It requires the depth knowledge and core concepts of the components and skills such as mathematics and statistics, programming and computing, machine learning and artificial intelligence, data visualization, domain knowledge in finance, healthcare, and e-commerce. The concept of ingestion, storage and processing, analysis and modelling, visualizing results and delivering actionable reports to the interested people and stakeholders are essential including the knowledge of python, R etc for the data science, data analytics, Machine Learning, Pandas, NumPy, and data engineering.

Keywords: (*Data Science, interdisciplinary, programming, ingestion, e-commerce, visualizing, python, ML, NumPy*).

INTRODUCTION

Python is a computer programming language. In Netherlands by a Dutch programmer, Mr. Guido Van Rossum was the first, who is writing the implementation of python in December 1989 but it was coming into the public and released in 20th February 1991 the first version of python 0.9.0. language. Mr. Guido Van Rossum is the successor to the ABC programming language. This programming language was named as per the name of British comedy series “Monty Python’s flying in Circus. As per October 2026, the current version of python is 3.15.0 has been released in the market. Many important features of the current version of python have made particularly for the useful in Data Science. Python offers numerous libraries that simplify compound logical household tasks. Such as Pandas cares data manipulation, NumPy runs numerical computing abilities, Matplotlib benefits generate graphs, and Scikit-learn propositions tools for machine learning in data science.

The computer programming language Python is cast-off for numerous arenas such as healthcare, education, agriculture, finance, environmental research, and business analytics etc. For example, many academic institutions can examine their student performance, health institute can study patient disease, and agricultural researchers can examine weather trends and agriculture production.

This study is a theoretical and descriptive study. This paper investigates the importance of Python in Data Science. Its purposes is to enlighten that why Python is extensively used and how its competences can be established with the help of a simple data analysis process in data science.

REVIEW of LITERATURE

Some literatures have been reviewed which include books, research publications, and technical documentation, web site, related to this topic for this study. The findings of this literature focus on Python programming, computing, data analysis, visualization and machine learning have been reviewed to find out the research gap.

Rossum. V. and Drake, (2003, 2006,2009,2011), “Python programming language various versions” clarifies its strategy and programming landscapes. Python highlights legibility and offers a flexible environment for emerging applications. These features make it appropriate for primary learners who want to acquire knowledge of python programming and data analysis processes. Python’s relatively straightforward syntax allows students to concentrate on solving problems rather than dealing with complicated programming structures.

Hunter, (2007), “Matplotlib as a flexible plotting library for creating visual representations of numerical data”. It explains graphs and diagrams support the scholar to understand data relations, circulations, and tendencies that may be problematic to classify after tables. Python visualization apparatuses are very much essential in academic and scientific experiments and statistical research.

McKinney, (2010), “introduced Pandas as an important tool for practical data analysis in Python”. Pandas delivers data constructions and purposes for establishing, filtering, converting, and analysing various forms of data. For example, students can use Pandas to calculate average marks, identify missing values, compare student performance, and summarize datasets.

Harris et al. (2020), “NumPy as a fundamental library for scientific computing in Python”. It explained multidimensional selections and scientific operations which supports’ well-organized numerical computation. It also explains that NumPy is useful as and when Data Science requires numerical computation concerning huge groups of mathematical values, statistical processes, and mathematical revolution.

[www.https://dl.acm.org](https://dl.acm.org), search on (23.10.2026), It explain the use of features of various versions of python useful in data science.

RESEARCH GAP

Although existing books, journal, research papers explain Python's libraries and their applications, but it may find it problematic to understand how these papers and books work together in a practical Data Science task with the active use of python computer programming language. Therefore, there find the existing research gap which is not yet covered properly.

RESEARCH QUESTIONS

This paper helps to find out the answer the following questions:

- Why is Python measured an significant programming language in Data Science?
- How do Pandas and NumPy aid in collecting, organizing, and analysing data in python library?
- How dose visualization libraries of python programming language helps the users to understand data used more efficiently?
- How can Python sustenance in decision-making in academic, agriculture, healthcare and other fields?

RESEARCH OBJECTIVES

The main objective of this study is to inspect the need of Python in Data Science and machine learning. The other secondary objectives are:

1. To elucidate the role of Python in Data Science.
2. To classify the needs of Python libraries applicable to for data science.
3. To validate the practice of Python in establishing and processing data in data science.
4. To define Python's advantages and limitation in data science.

RESEARCH METHODOLOGY

Research Design

The study adopts a descriptive and empirical approach with using the existing secondary documents, research papers, reports etc. which have already been published about python computer programming languages ‘and its application into the data science. The have also been collected from the internet through google search. Data may be collected and to be shown in a tabular form from the from our school at random for five students of class twelve class test for the subjects Physics, computer science and mathematics are shown in the following table:

Students no.	Physics	Computer Science	Mathematics
01	85	72	77
02	83	99	88
03	78	91	87
04	90	93	91
05	79	89	95

TOOLS AND TECHNIQUE USED IN THIS STUDY

In this study various tools like Python, Pandas, NumPy, Matplotlib etc. by using the creation of data set, load the data, statistics, visualize the results, Interpretation and observation a python programming is written.

SAMPLE PYTHON PROGRAM

A list of numbers from number 1 to number 10 can be written by the python are:

Numbers = [1,2,3,4,5,6,7,8,9,10] but #using this list comprehension to create a new list containing only even numbers

Even_numbers = [x for x in numbers if x % 2 == 0]

Print the new list of even numbers

Print (even _ numbers)

IMPORTANCE OF PYTHON IN DATA SCIENCE

Python is important in Data Science because it combines a readable programming language with a broad collection of tools for processing, analysing, and interpreting data.

- **Simplicity and Readability:** Python uses clear, elegant and concise syntax that makes it easy to learn, understand and read, even for people new to programming.
- **Flexibility:** Python can be used for a wide variety of tasks such as web development, software development, scientific computing, data analysis, etc.
- **Large Open-Source Library:** Python has a large open-source library that provides various useful functions and modules.
- **Community Support:** Python has a large and active community of programmers who continuously contribute to its libraries and other resources, making it easier to work with the language.
- **Multi-Platform Support:** Python can be used on many different operating systems such as Windows, Mac OS, Linux, etc.
- **Interpreted:** Python is an interpreted language, meaning an interpreter executes it rather than being compiled into machine code. This enables a faster development life-cycle.
- **Object-Oriented:** Python supports object-oriented programming.

MACHINE LEARNING

Machine learning is a division of Artificial Intelligence that empowers computer systems to learn designs from data and build calculations and classifications data required in data science. Python is extensively used for emerging and assessing machine learning system required in data science. Scikit-learn consist of classification of data, regression model, clustering process, and model assessment in data science process. In other word, the Machine learning is the subset of Artificial intelligence that gives lesson that how computer predict from data and take decision without any instruction However, a relationship between two variables does not necessarily prove that one causes the other. Predictions also depend on data quality and the suitability of the model.

ADVANTAGES OF PYTHON IN DATA SCIENCE

The main advantages of python are as follows:

1. **Easy to learn:** It is very much readable syntax and makes it suitable for beginners of python language user.
2. **Extensive libraries:** python provides important tools for mathematical computing, statistics, visualization, and machine learning which helps in data science.
3. **Open-source software:** Python and its popular libraries can be used without buying a programming-language licence from the producer.
4. **Versatility for supports:** Python generally supports in many jobs which varying from simple type of computation to advanced type of machine learning.
5. **Support to the Community:** A large volume of community engaged in tuition, record keeping, documentation and problem-solving activities for common programming problems.
6. **Mechanization:** It can perform efficiently the repetitive and analytical type of operations.
7. **Cross-platform support:** Python is very easy and available for application on most of the operating systems.
8. **Composite value:** python helps the learner in combine programming for many streams like computer mathematics, statistics, and logical reasoning type of problems.
9. **Tools Used:** python can easily use various type of tools and techniques.

DISADVANTAGES OF PYTHON

Python computer programming language has the following limitations. These are:

Lower execution speed: in Some cases, Python language operations have slower than corresponding implementations in accumulated computer languages. Sometimes numerical libraries used in python often discourse this with the help of optimized code written in lower-level languages in python language.

Huge Memory consumption: Generally Large volume datasets may require considerable memory, particularly when several numbers of dataset have been created.

Doubtful Data quality: Python language cannot inevitably guarantee that the source data is correct, comprehensive, or illustrative in all cases.

Model reliability: in case of inadequate training data, Machine learning predictions may be inaccurate and inappropriate. Hence, Python should not be given a guarantee of correct analysis.

CONCLUSION

This study observes the importance of Python in Data Science through a evaluation of existing literature by using the secondary data sources.

This paper elucidates that how Python language supports various data collection, cleaning, organization, statistical analysis, visualization, automation, and machine learning process. a guarantee of correct analysis. Tools used in Libraries such as Pandas, NumPy, Matplotlib, and Scikit-learn shorten diverse phase of the Data Science procedure. It actually allows various python users to grow logical answers in the area of a common python language programming atmosphere.

This has been observed that Python can calculate averages of data and generate empirically statistics from a structured dataset. These competences help to transform basic numerical data into meaningful insights. This has supported the interpretation and decision-making in data science.

The main strengths of python programming language include its legible syntax, wide-ranging libraries, open-source ecosystem, versatility, and strong academic value. Its disadvantage includes possible presentation and memory constraints, dependence on data quality, and the necessity for suitable mathematical and statistical data in python

The study may be concluded that Python is an important programming language for learning and for applying into data science process.

This paper could compare Python language with another computer programming language, examine its efficiency on greater datasets, or observe its utility in analysing actual enlightening or ecological data through python.

BIBLIOGRAPHY

1. **Rossum. V. and Drake**, (2003), "Python programming language various versions" Python Reference manual, CreateSpace.
2. **Rossum. V. and Drake**, (2006), "An Algorithm for Reconstruction of Entrance Beam Fluence from virtual patient Exit Electronic Portal Image" IJMPER.
3. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.
4. **McKinney**, (2010), "Data structures for Statistical Computing in Python" 9th Python Science congress (SciPy 2010), Austin, Texas.
5. **Pedregosa, F., Varoquaux, G., Gramfort, A., et al.** (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
6. **Harris et al.** (2020), "Array Programming with NumPy", Science journal Nature, Vol 585, Pages 357-362.
7. **www.https://dl.acm.org**, search on (23.10.2026), It explain the use of features of various versions of python useful in data science.
8. Python Software Foundation. *Python Documentation*. Available at: <https://docs.python.org/3/>
9. The pandas development team. *pandas Documentation*. Available at: <https://pandas.pydata.org/docs/>
10. NumPy Developers. *NumPy Documentation*. Available at: <https://numpy.org/doc/>