

# Implementing Data Visualisation Using Various Python Libraries

Pragati<sup>1</sup> Minakshi Arora<sup>2</sup>

<sup>1</sup>M.Tech. Scholar <sup>2</sup>Assistant Professor

<sup>1,2</sup>Department of Computer Science and Engineering

<sup>1,2</sup>SKITM, Bahadurgarh, India

**Abstract** — Today's organizations must perform data analysis and visualization. The field of data science deals with the preparation, cleaning, and analysis of data. Numerous industries, including travel, healthcare, gaming, financial services, digital advertisements, and internet searches, use data science algorithms. Today's environment requires data analysis and visualization for a wide range of applications. Classification, anomaly detection, quantification, organizing, and decision-making challenges are among the concerns that data science can resolve. We have demonstrated in this study how Python is helpful in solving these kinds of difficulties. Apart from Python, there exist alternative platforms that are employed to address tasks that are exclusively grounded in data science. Here, we've concentrated on Python and its libraries, which are quite helpful for issues with data science. We've demonstrated the usage of Python in data analysis and visualization.

**Keywords:** Data Science, Python, Data analysis, Data Visualization

## I. INTRODUCTION

Because data science can use the data itself to solve decision-making difficulties, it is a very appealing field. It goes considerably beyond this, in actuality. It typically uses regressions and classifications to solve problems. Classification-based tasks include determining the result from a set of fixed values. Regression-based issues, such as Monday's temperature forecast. Problems based on clustering to comprehend data structure. For example, to determine which viewers enjoy the same film.

The aforementioned issues are all industrial in nature and are incredibly beneficial for any firm looking to move forward and establish a solid position for itself.

Data science provides thoughtful answers to challenging queries. Data science can be used to solve a problem such as determining the stock's sale price for the upcoming week for a company. Uncertainty in question-answering may not be helped by data science. Questions should be specific enough to have facts to support their answers. The entirety of data science is not applicable to one individual. The group includes specialists in various pertinent domains, computer scientists, statisticians, and artificial intelligence scientists. By definition, data science is multidisciplinary and necessitates both domain science and technique disciplines [1]. Numerous fields are included in data science, including mathematics, statistics, information theory, machine learning, and optimization. Depending on the discipline, different data may be used in the procedure. Bioinformatics, social informatics, and other fields employ scientific data.

However, in the event that company data is available, organizations such as Amazon, eBay, Google, Facebook, and others utilize it to develop or forecast new commercial strategies. Amazon use collaborative filtering to

suggest premium products to its customers, whereas Facebook utilizes the "people you may know" tool to suggest friends. We'll talk about how Python can be used to solve each of the data science problems listed in this paper. For data science, a wide range of tools and technologies are employed. Python is one of the most popular and useful programming languages for data research. It serves as the data science key. Python has numerous features that make it a great tool for data science. We'll talk about the intriguing features and libraries—like pandas, numpy, matplotlib, and others—that make Python an extremely practical and efficient tool for data research. We will also talk about the features that are relevant to data science and involve Python at all skill levels.

## II. STEPS IN DATA SCIENCE

The following are the steps involved in a data science process:

- Acquire data
- Preparation of data
- Analysis of data
- Preparation of report
- Apply results or Act on results



Fig. 1: Steps involved in data science process

- 1) Acquire data: It entails locating data sets and retrieving the information needed to solve the issue. For good outcomes, the data must be accurate. Data can be gathered in a variety of methods, such as by searching online or by gathering it from the outside world. However, it must be pertinent to the issue we are attempting to resolve (Determine data sets, Get data, Ask for data).
- 2) Prepare data: After data collection, the information is ready for analysis. The most crucial action to take is this

one. We receive better results the more time we spend on this section. In order to prepare the data, we first investigate it through a preliminary study to determine its nature. Next, we clean and integrate the data as part of the preprocessing step. Preprocessing entails resolving any issues that arise from the imperfect data, such as redundant information, missing values, and unusable features. For example, missing values might be filled in based on features.

It could be the feature's mode, the average of all accessible values, or something else entirely. From the perspective of outcomes, therefore, this data science stage is crucial. It takes better preprocessing to achieve better results. Working with relational and labeled datasets is made simple with the help of the Pandas package, which offers quick and adaptable data structures [2]. Given that 70% of the time required to make the data set usable and compliant with algorithm requirements is spent preparing the data set. Selecting cardinal features, deleting unnecessary information, and lastly converting raw data into sessions are the objectives of data preprocessing.

- 3) **Analyze data:** (Choose analytical approaches, Create models): After the data is ready, an analytical technique is applied to it in order to create a model. A model is constructed to allow for the graphical visualization of data and, crucially, its analysis. Analysis entails searching for the outcome based on specific characteristics.
- 4) **Preparation of report:** After the results are analyzed, a report is created using the features that are pertinent to the issue at hand. Decisions or actions for the future point of view might be made with the help of this analytical report.
- 5) **Act:** As soon as the report is finished, something happens. While a definitive answer is obtained, the pertinent problem is taken into consideration while making decisions. The outcome could take the form of reinforcement, clustering, classification, or another method such that analysis will be used to inform decisions in the future. Iterations make up the entire procedure.

### III. USE OF PYTHON

The Python programming language has several applications in data science. Python can be used to tackle data science activities such as data analysis and visualization, which are essential components of data science. Python has a large number of libraries that can be used to accomplish this kind of work. In actuality, the Python programming language may be used for everything from data preparation to data visualization. We'll talk about a few libraries and their purposes here.

Since data analysis and data visualization are key to data science. Great libraries exist in Python for these kinds of jobs. To examine the data using different numerical data analysis approaches, one might utilize the NumPy library. SciPy is an additional library that is useful for statistical analyses. Similarly, graphs for the analysis can be created using the Matplotlib tool. As is well known, the creation of

an effective model depends greatly on data preparation. Additionally, Python is utilized to prepare the data.

Python can readily handle such jobs since, as we have seen, cleansing, integration, and other related operations are components of preparation.

As we can see, the Python programming language may be used for all activities, from data preparation to data display. In software such as ModelBuilder, scripts written in the Python programming language are more advantageous than models. It is possible to use simple program structures like loops (for, while), decision-making statements (if, else), etc. Additionally, there's a good chance that errors won't occur while the software runs [5].

Another pandas library that is compatible with Python is available. This is the rich data structure library and toolkit that is commonly used in the social sciences, finance, statistics, and many other sectors to work with structured data sets. The key issues are feature selection and extraction, aside from these helpful libraries [6].

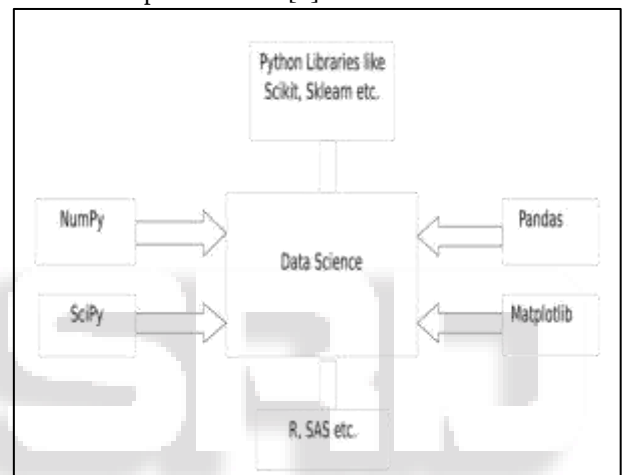


Fig. 2: Python Libraries with other skills required for data science

MongoDB database can be used to store the dataset that will be used for applying data science processes. So the above discussed libraries are as follows: Pandas (for data manipulation and analysis)

- NumPy (for analysis)
- SciPy (for statistical tests)
- Matplotlib (for graphical presentation)

There are numerous methods in these libraries that are utilized to do particular tasks. Thus, these libraries are really helpful and strong for Python data science tasks. Other kinds of libraries are also available for Python that can be used for various computing tasks. For example, the multiprocessing Python library is a package that facilitates the spawning of subprocesses in order to take advantage of multicore shared-memory capabilities [4].

### IV. MACHINE LEARNING USING PYTHON

The field of machine learning deals with the creation of algorithms based on data. Data is made up of testing and training data. The model is prepared using training data, and its correctness or performance is verified by applying testing data to the model. Various machine learning methods exist, including reinforcement, unsupervised, and supervised algorithms. Regression and classification are performed

using supervised algorithms. For clustering, unsupervised methods are employed. We obtain a result when testing data is applied to the model that was created using machine learning techniques and training data. To offer the analysis in its ultimate form, these results are evaluated and visualized. Businesses and organizations use visualizations to help them make decisions that will maximize their benefits overall.

#### V. SKILLS REQUIRED FOR DATA SCIENCE

Python's significance for data science is evident. Preprocessing, analysis, statistical testing, visualization, and other uses are all possible with it. Apart from proficiency in Python, a data scientist should possess additional talents. Working with unstructured data should be a skill for a data scientist.

Generally, a developer can work with organized data with ease using a SQL database; however, the Hadoop platform may be utilized to work with unstructured data. Data scientists need to be proficient in math and analysis, among other subjects. A data scientist should be familiar with SAS and R in addition to these. But no organization's data science work is completed by a single individual. Although there is a dedicated team for that, being a data scientist requires having the aforementioned abilities. Analyzing the situation reveals that it functions similarly to artificial intelligence in that data is used to make predictions. The intelligence of a (fictitious) machine that is capable of effectively completing any intellectual work that a human being can is known as human-level artificial intelligence [7].

#### VI. CONCLUSION

The data science process, which includes gathering raw data, preparing it for analysis, communicating the findings, and taking appropriate action, has been demonstrated to us. Thus, the Python programming language is quite effective and helpful for both data collection and data visualization. Python libraries for various purposes have been covered, including Numpy for analysis, Pandas for data manipulation and analysis, Scipy for statistical tests, and Matplotlib for graphical result presentation. There are numerous other Python libraries that are useful for data science in addition to these ones. Additional languages and tools that serve diverse purposes are also available.

As we previously covered, data science also makes use of SAS and R. Preprocessing data can be done with the R language, although Python can also be utilized. As a result, it is evident that the Python programming language is a viable option for carrying out numerous data science-related tasks. Python serves as a crucial programming language for data science, so to speak. We can draw the conclusion that Python is a versatile language that offers useful tools for carrying out data science procedures and executing programs [8].

#### REFERENCES

- [1] Javin D. West, "The science of data science", Journal of Integrated creative studies, No. 2016-010-e, May 2022.
- [2] Wes McKinney, "pandas: a Foundational Python Library for DataAnalysis and Statistics", DLR Portal,

[www.dlr.de/sc/Portaldat/15/Resources/dokumente/.../p\\_yhpc2011\\_submission\\_9.pdf](http://www.dlr.de/sc/Portaldat/15/Resources/dokumente/.../p_yhpc2011_submission_9.pdf).

- [3] Gaurav, Zunaid Alam, "Road Safety in india using Data Mining Approach", International Conference, REDSET 2017, CCIS 799, [https://doi.org/10.1007/978-981-10-8527-7\\_17](https://doi.org/10.1007/978-981-10-8527-7_17)
- [4] Rosa Filguiera, Iraklis Klampanos, Alexander Moreno and Malcolm Atkinson, "A Python Framework for Data-Intensive Scientific Computing", IEEE Conference, 978-1-4673-6750-9, Nov-2018.
- [5] Ing. Zdena Dobesova, "Programming Language Python for Data Processing", IEEE Conference, 978-1-4244-8165-1/11, Sept-2021.
- [6] Fabien Dubosson , Stefano Bromuri, and Michael Schumacher, "A Python Framework for Exhaustive Machine Learning Algorithms and Features Evaluations ", IEEE Conference, 1550-445X/16, March-2019.
- [7] Ankur Bhatia, "Artificial Intelligence – Making an Intelligent personal assistant", International Journal of Computer Science and Engineering, Vol. 6, No. 6, 2019.
- [8] Nurul Afiah Mat Zaib , Nor Erne Nazira Bazin , Noorfa Haszlinna Mustaffa , Roselina Sallehuddin, "Integration of System Dynamics with Big Data Using Python: An Overview", IEEE Conference, 978-1-5090-6255-3/17, May 2020.