Chapter 6
Exploratory Data Analysis in Python

R. Sruthi
Coimbatore Institute of Technology, India

G. B. Anuvarshini
Coimbatore Institute of Technology, India

M. Sujithra
Coimbatore Institute of Technology, India

ABSTRACT

Data science is extremely important because of the immense value of data. Python provides extensive library support for data science and analytics, which has functions, tools, and methods to manage and analyze data. Python Libraries are used for exploratory data analysis. Libraries in Python such as Numpy, Pandas, Matplotlib, SciPy, etc. are used for the same. Data visualization's major objective is to make it simpler to spot patterns, trends, and outliers in big data sets. One of the processes in the data science process is data visualization, which asserts that after data has been gathered, processed, and modelled, it must be represented to draw conclusions. As a result, it is crucial to have systems in place for managing and regulating the quality of corporate data, metadata, and data sources. So, this chapter focuses on the libraries used in Python, their properties, functions, how few data structures are related to them, and a detailed explanation about their purpose serving as a better foundation for learning them.

DOI: 10.4018/978-1-6684-7100-5.ch006

INTRODUCTION

The vital process of performing primary analyses on data to explore new trends, patterns, identify outliers, hypotheses, and cross-check assumptions using summary of data's statistics and graphical representations is known as exploratory data analysis. Exploratory Data Analysis (EDA) is not a formal procedure with rigid guidelines. The approach to EDA emphasizes open-mindedness and the exploration of various notions in the early stages of analysis. Some concepts utilized may succeed while others may fail, highlighting the beauty of EDA as a means of exploring uncertainty. Despite the availability of predefined methods and techniques, exploratory data analysis remains a crucial component in any data analysis. As research progresses, certain fruitful areas will be discovered, documented, and shared with others.

PYTHON LIBRARIES TO PERFORM EDA

Numpy

NumPy is a powerful Python library for efficient array manipulation and numerical computing (Harris et al., 2022). Originally known as "Numeric," it evolved into NumPy with enhanced capabilities. Its optimized C programming and extensive functions make it the standard for array computation, supporting tasks like linear algebra and Fourier transformations. With a vibrant community and easy accessibility, NumPy is widely used in data analysis, machine learning, and more.

Pandas

Pandas is a Python library that efficiently manipulates diverse data, collaborating seamlessly with other data science libraries like NumPy, Matplotlib, SciPy, and scikit-learn. It plays a vital role in machine learning by accurately exploring, cleaning, transforming, and visualizing data. Jupyter notebook facilitates easy execution of Pandas programs, enabling data visualization and analysis. Pandas expand Python's data analysis capabilities with essential procedures: data loading, manipulation, preparation, modeling, and analysis (Ateeq & Afzal, 2023).

SciPy

SciPy is a powerful Python library that builds upon NumPy, providing multidimensional arrays for scientific and mathematical problem-solving (Khandare et al., 2023). It eliminates the need for separate NumPy imports and is widely used in Machine