ISSN 2278-3091
Volume 10, No.1, January - February 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse391012021.pdf
https://doi.org/10.30534/ijatcse/2021/391012021

Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python

Ali Hassan Sial [1], Syed Yahya Shah Rashdi[2], Dr. Abdul Hafeez Khan [3]
[1]Department of Computer Science, SMI University, Karachi, Pakistan, sial_alihassan@yahoo.com
[2]Department of Computer Science, SMI University, Karachi, Pakistan, yahyarashdi6@gmail.com
[3]Department of Software Engineering, SMI University, Karachi, Pakistan, ahkhan@smiu.edu.pk

ABSTRACT

With the tremendous growth in the areas of computing, statistics, and mathematics has led to the rise of the emerging field of expertise, named 'Data Science'. This paper focuses on the comparative study and evaluation of the data science libraries used in Python Programming Languages, named 'Matplotlib' and 'Seaborn'. The sole purpose of this paper is to identify areas and evaluate the strengths and weaknesses of these libraries with the implementation of code and identify the classification of the univariate and multivariate plotting of data concerned with patterns of data visualization and computational modelling of data in the form of processed information using techniques of big data and data mining.

Key words : Data Visualization, Computational Modelling, Univariate, Multivariate, Big Data

1. INTRODUCTION

Data Visualization is the graphical illustration for a pictorial representation of data with the integrated use of illustrated design. The sole perspective is to provide in a visualized form that is easier to understand and presented. In a general perspective, data visualization techniques can be classified into two categories 1) univariate and 2) multivariate data visualizations. The first category, univariate data visualization constitutes of plotting a specific variable to identify and understand relatively more about the distribution and scattering of plots whereas, multivariate plots classifies the relationship of several datasets and variables [2, 3]. The popular data visualization techniques that comprise of scatter plots, bar charts, pie charts, and line charts, are extensively used in the areas of data science, mathematical modelling, and computational research. In a greater extent of the rapid transformation of data or information, although new techniques are used to visualize quantitative and qualitative information for data researchers to incorporate data analytics and mathematical computations for better efficiency and performance metrics. The tools of data mining are broadly categorized into three types 1) programming languages 2) Statistical tools 3) visualization tools. The merits of SPSS and Stata both fall under the category of statistical analysis software packages that are solely used for the management or organization of the datasets. The researchers have identified that SPSS in various data visualizations areas of complicated and complex data analysis alongside Stata can be utilized for high-level areas in the research and development industry. Furthermore, R is a high-level, resource-oriented data analysis package and high-level programming language that is used for numerical and statistical analysis, data visualization and reporting. Significantly, Python is a powerful, high-level, general-purpose, and resource intensive programming language. Thus, the key difference among R and Python is that R is a statistical, numerical and data analysis-driven programming language whereas Python is a general-purpose programming language.

A. *Matplotlib*

Matplotlib is one of the most popularly used data visualization libraries of python. This library was built by a John Hunter who is along with several contributors, and it had put in a greater amount of time into prompting this software used by every scientist and philosopher across the globe [4]. Matplotlib is a graphics library for data visualization package in Python which encompasses as an integral aspect in the python data science stack and it is easily supported with NumPy, Pandas and other relevant libraries.

B. *Seaborn*

Seaborn is a graphic visualization library that is built on the primary configurations of Matplotlib. It provides accessibility to the users with some of the most commonly provides data visualizations processes with certain data visualizations necessities such as mapping colour to a variable or using faceting requirements across the globe. It provides seaborn is more integrated for working with Pandas DataFrames [4].

C. *Pandas*

Pandas is an open-source library used in Python that provides enhanced performance metrics, easy to use data structures and data analysis packages, tools and libraries for Python Programming Language [4][5][6]. The use of pandas with

python encompasses various fields of expertise including data sciences, computational modelling, finance, economics, statistical analysis, machine learning, etc.

D. *NumPy*

Numerical Python (NumPy) is a data analysis and numerical computation library that consists of multidimensional objects of the array and a collection of routines or procedures that are used for the processing of similar types of arrays. Using the NumPy library, data scientists, system programmers and architects can perform mathematical computations, numerical and logical operations on these arrays to build a sustainable and efficient computational model for scientific and numerical perspective.

2. RELATED WORK

Psallidas, Fotis, *et al.*, presented an overview of data sciences tools and techniques including NumPy, TensorFlow, and PyTorch, that represents the comprehensive repositories as well including GitHub and vice versa constitutes of machine learning principles and techniques to promote latest applications and updates [1]. Stančin, I, and Jović, A, discussed about the emerging trends and growth expansion of big data, and data mining techniques and tools with an overview and comparative analysis of the python libraries that focused on data preparation, data visualization, deep learning, machine learning, and information retrieval. The authors compared Matplotlib, Plotly, Scikit learn, TensorFlow, Keras, PyTorch, Hadoop Streaming and PySpark [2]. Odegua, and Ikpotokin presented a concept of a python-based library for data visualization, big data analysis, and computational modelling, named as "DataSist", which provides high-level, flexible, and easy to use methods, and functions that can provide assistance to data scientists and big data analysts to rapidly analyse, interpret and visualize large data sets in a manageable and intuitive form [3]. Oberoi, A. and Chahuhan, R. expressed a detailed overview of comparison between seaborn and matplotlib on the basis of univariate such as line plots, bar plots, histogram, box plots and multivariate plots such as scatter plots, regplots, joint plots, pair plots, and heat maps, the authors has comparatively analyzed matplotlib and seaborn using various tools such as Numpy, and Pandas DataFrames for specific design and performance metrics to promote the notion of data sciences [4]. Barret, P, *et al.*, expressed the concept of matploltlib library, it is a portable 2D plotting package built in Python. The primary focus of matplotlib is to perform graphical visualization of scientific, engineering, economical & financial data in pictorial illustration. The authors have also identified the features of matplotlib, and its working procedure. The major aim of this library is to provide an integrated package with improved performance metrics and features that are useful for scientists, architects, programmers & engineers [5]. Fahad, A, SK, and Yahya, E, Abdulsamad, discussed a review on "Big Data Visualization using R and Python with GUI Tools", the authors showcased the notion of "Data Visualization" and its implementation using high-level programming languages, Python & R with the persistent information that can be used to assist to cultivate provocations and assistance to aid futuristic support and technical capabilities. Moreover, Data Visualization Techniques are classified and authenticated in a scientific form to surpass thousands of times that are reliable instead of the textual representation of the raw data into processed information [6]. Gaurav, and Sindhu, Ritu expressed the conceptualization, technicalities, and understanding of "Data Science Concepts, Tools & Techniques in Python", the algorithms & methodological interpretation of these techniques in the quantitative form to quantify and find out a way to organize, make decisions and solve problems using multidimensional libraries in python. The authors have presented a review on the principles of data science using NumPy, SciPy, Pandas, and Matplotlib for graphical, numerical, and statistical analysis of datasets in a step-by-step form of data analysis, interpretation and processing of raw data into a manageable form (graphical & statistical measures of the predicated results) [7]. Data Analysis & Statistical Measures were compiled and contrasted in a comprehensive guide which highlighted the key parameters of python libraries, specifically seaborn and matplotlib. The overall compilation and process of working with python libraries for data visualization constituted a range of steps to further explore and identify the datasets, plot and visualize them with numerical calculations and critically analyse the parameters of box plots, bar charts, line plots, and histograms. Furthermore, the authors have compared and contrasted matplotlib and seaborn based on mapping data, 2D visualization & improved performance metrics [8].

3. DATA ANALYSIS

There are two types of data first one is structured that is small data which is in MB, GB, TB second one is unstructured data that is BIG DATA which is in PB, EB.90% of the world has now only big data that is unstructured data unlike RDBMS for example Facebook and google generates more than 5 petabytes of data daily. Velocity of big data is very fast it is increasing exponentially. Storage has been done in distributed form and globally available. Hadoop, Spark are generally softwares used to deal with big data. Multi-node cluster are used for traffic handling in big data. Significantly, the techniques of data analysis includes Data Mining, Business Intelligence, and Descriptive Statistics. There are briefly explained in the sections below:

A. *Data Mining*

Data mining is a specific kinds of data analysis in which we extract useful information from bulk of data. Knowledge Data Discovery is also term used for data mining in which we use steps for extraction of data, data cleaning, integration, selection, transformation, data mining, knowledge evaluation and representation.

B. *Business Intelligence (BI)*

Business intelligence is the process in which we use our planning and technology to analyze data and information of organization, to achieve certain goals and to take certain decisions and planning and make strategies to gain profit in the form of statistical and computational applications for data analysis may be categorized in the form of descriptive statistics, prospective data analysis, predictive analysis and experimental data analysis.

4. DESCRIPTIVE STATISTICS

Descriptive Statistics presents a comprehensive findings about observations, results or sampling of data or information. The collected statistical analysis can be considered as quantitative, statistical summaries such as mean, mode, median, percentiles, maximum and minimum ratio, visualization charts such as graphs and plots. The overall analysis and evaluation of descriptive statistics that may be distributed into univariate, bivariate and multivariate analysis of the statistical measures. The further exploration and explained below:

A. *Univariate Analysis*

Univariate analysis is considered as the foremost ways for the explanation for data or information. This involves a single variable and doesn't focus with reasons or liaison and the key persistence of univariate is to describe among the mentioned instances.

B. *Bivariate Analysis*

Bivariate analysis is the investigation of two dimensions that are compared alongside to identify and potential relationships amongst themselves. It deals with causes and relationships and the major purpose of Bi variate data is to explain. Bi variate data uses analysis of two variables simultaneously.

C. *Multivariate Analysis*

In this method data can be depend upon 3 or more variables within 200 variables.it helps to summarize the data. It also reduce the chance of data.

5. DATA PROCESSING & COMPARISION

Data parsing and identification is done and then correction is done by using algorithms available, standardization is done and measurements are standardized, matching are used for removing duplication and space is freed from duplication of data, consolidation of the records are done and records are matched again, Missing data is searched and incorrect data is deleted.

A. *Data Cleaning*

Data cleaning is classified as an aspect of the data processing segment. Seemingly, data cleaning arises after the data or information has been processed and planned in them an age able form. In this phase, the received data or information might comprises of the replacements, duplicated features, bugs, errors or be inadequate. This refers to the needs of data or information that is needed for cleaning of data or usage of information. Some of the common task done in this stage may include data replication, normalization of datasets, segmentation of data features, characteristics of record matching etc.

B. *Data Visualization*

Data visualization is the graphical representation of data. It constitutes of all the relevant processes and techniques that were involved in the communication mediums of data or information in illustrative form that seems easy to identify, examine and respond to the queries. "A picture is worth a thousand words" human brain can easily understand images and pictorial form of data from numbers and figures because it is easy to capture images that are in graphical form. From graphical form of data we can easily interpret results and take decisions from data, we can use different variables in the data to show effects. There are steps for data visualization first step is to transform data set secondly we will find insights then analyze the data and finally we will visualize the data. Matplotlib is package in python used for plotting 2D graphs e.g. bar chart, line chart, pie chart, scatter plot, histograms, hexagonal binplot, and area plot as shown in the Figure1.1[2].

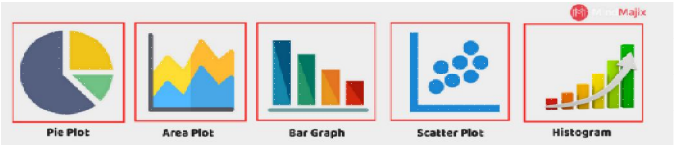

Figure 1.1. Graphical Representation of Data

Bar graph is used to compare one or two categories such as male female, two regions such as east and west and have y and x axis. Histograms are used for data distribution charts by making intervals e.g. by making distribution of customers of ages 5- 10,11-15,15-20. scatterplot is used for comparing two measures i.e. profit and sale for many companies we will compare by using x and y axis and put sales and profit on different axis and compare using dots, hence dots are used represent profit and loss, pie chart are circular graphs that shows composition between different regions using slices of pie, hexagonal chart shows the values using density and is best for population.

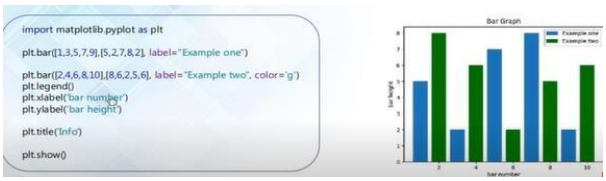

Figure 1.2. Bar Graph Representation using Matplotlib

The process of data visualization in matplotlib and seaborn distinguishes in the form of manipulating the datasets, and visualizing them in bar charts, pie plots, and scatter plots, line plots, classifying them into multivariate and univariate analysis. The following illustrations shows the difference between matplotlib and seaborn in the form of data visualization, statistical analysis and performance metrics.

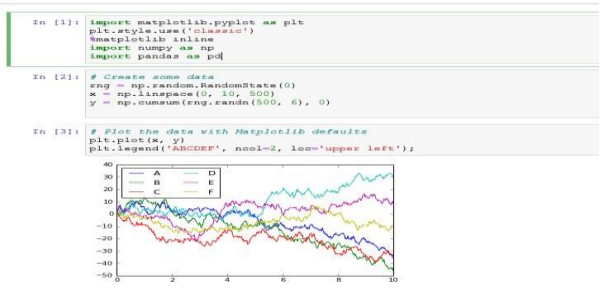

Figure 1.3. Graphical Illustration using Matplotlib, Numpy and Pandas

The following example shows that after incorporating the seaborn library, the overall code implementation and data visualization patterns become interactive and well-illustrated, this is because of manipulation of data using various datasets.

Figure 1.4. Graphical Illustration using Matplotlib, Seaborn, Numpy, and Pandas

The example uses the 'flights' dataset and plots a scatter plots that shows the number of passengers in the form of data and month as shown in the figure 1.3. This feature extracts the overall information using data visualization libraries i.e. matplotlib, numpy, pandas and seaborn.

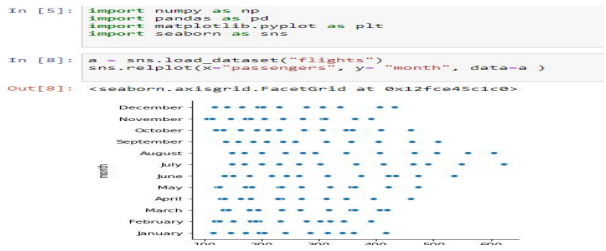

Figure 1.3. Illustration of Scatter Plot using Datasets

Scatterplots are used to represent the impact of one variable is affected by another variable, the relationship within two variables is considered as correlation.The close intensity of the data points that arises when they are plotted to draw a straight line, the higher or closer the correlation among the two variables or the strengthened association. The example below shows the scatterplot representation using matplotlib in python.

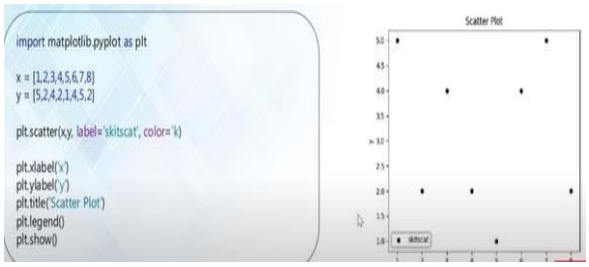

Figure 1.4. Visualization of Scatterplot using Matplotib

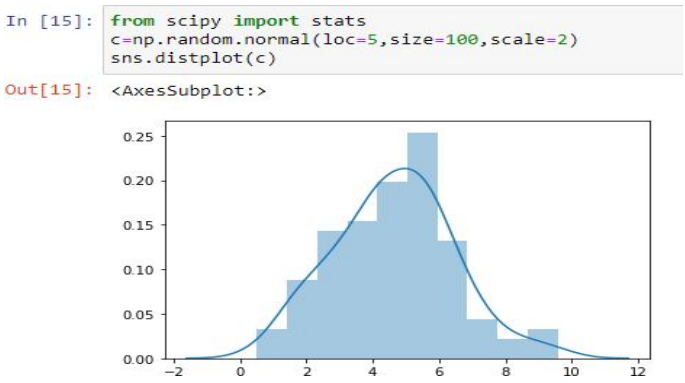

Figure 1.5. Visualization of Univariate Distance using Displot Function

The above histogram shows the univariate distance by passing of the curve using Displot Function using scipy and seaborn data visualizations libraries comprises of numerical functions that predicts graphical representation in the form of datasets and numerical computations.

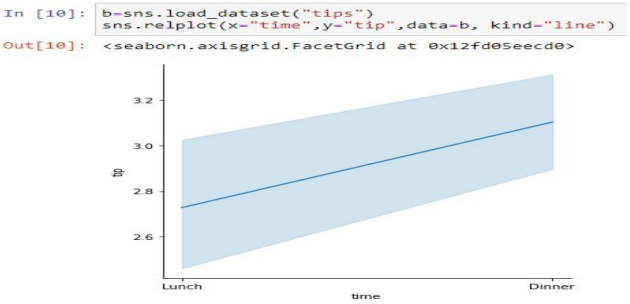

Figure 1.6 .Visualization of Line Plot using Dataset in Seaborn

The above graphical illustration shows the Line Plot representation of the dataset "tips" as shown in the figure 1.6. [4], the use of univariate and multivariate variables in this dataset are distributed in the form of systematic code and provides a classification of the numerical and statistical computations as represented in the graphical representation.

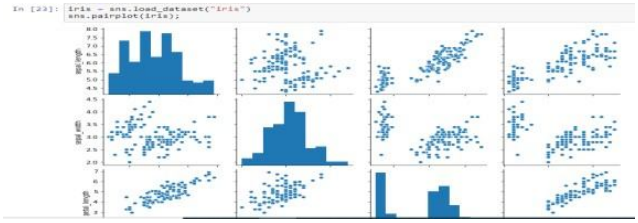

Figure 1.7. Visualization of Iris Dataset

The above Iris Dataset comprises of the all the graphical visualization patterns constituting variables present in both univariate and multivariate plots represents the data in visualized form i.e. barplots, scatterplots, histogram,lineplots and distributed plots as shown in the figure 1.7. [5]. Furthermore, the patterns of data are affected by the manipulation of data from the libraries and the processed information is transformed in mathematical methods and numerical analysis is conducted in graphical representation of datasets.

6. CONCLUSION

In this paper, data visualization libraries are discussed and evaluated using Python. The overall analysis and examining of the principles and concepts presents a detailed review about analyzing the patterns of numerical variables and datasets that processed and executed using Python. The comparison of seaborn and matplotlib, explains the code structure, functionality and performance metrics of both the libraries, additionally we have also incorporated alternative libraries to execute the source codes in a manageable and organized manner. It has been identified that if a data scientist wants to visualize the large chunks of datasets then seaborn will be a better option, but if you are looking for basic visualization patterns then matplotlib would be a better choice for beginners and starters in the field of data visualization & computational modelling. Furthermore, this paper will be a reliable source for starters and professionals to understand and examine the role of data visualization libraries in the various areas of expertise, i.e. Data Mining, BigData, Computational Statistics and Numerical & Symbolic Computation.

REFERENCES

[1] F. Psallidas *et al.*, "Data Science through the looking glass and what we found there,"2019.

[2] I. Stancin and A. Jovic, "An overview and comparison of free Python libraries for data mining and big data analysis," *2019 42nd Int. Conv. Inf. Commun.Technol. Electron. Microelectron. MIPRO 2019 - Proc.*, pp. 977–982, 2019, doi:10.23919/MIPRO.2019.8757088.

[3] A. Oberoi and R. Chauhan, "Visualizing data using Matplotlib and Seaborn libraries in Python for data science," *Int. J. Sci. Res. Publ.*, vol. 9, no. 3, p. p8733, 2019, doi:10.29322/ijsrp.9.03.2019.p8733.

[4] F..O.I.Rising.O.Odegua,"DataSist:APython-based library for easy data analysis, visualization and modeling," 2017.

[5] P. Barrett, J. Hunter, J. T. Miller, J.-C. Hsu, and P. Greenfield, "matplotlib -- A Portable Python Plotting Package," *ASP Conf. Ser.*, vol. 347, no. June, p. 91, 2015.

[6] S. K. A. Fahad and A. E. Yahya, "Big Data Visualization: Allotting by R and Python with GUI Tools," *2018 Int. Conf. Smart Comput. Electron. Enterp. ICSCEE 2018*, no. July, 2018, doi: 10.1109/ICSCEE.2018.8538413.

[7] .G.andR.Sindhu,"Python as a key for Data Science," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 4, pp. 325–328, 2018, doi:10.26438/ijcse/v6i4.325328.

[8] O. Sehgal, "Visualizing data using Lattice in R and Seaborn in Python for data science," *Int. J. Sci. Res. Publ.*, vol. 9, no. 12, p. p9609, 2019, doi: 10.29322/ijsrp.9.12.2019.p9609.