Methods and Tools for Data Analysis and Visualization

Vinod Kumar Yadav
Phd Research Scholar
Computer Science and Engineering Department
Rabindranath Tagore University Bhopal
vinod.it210@gmail.com

Dr. Pritaj Yadav
Associate Professor
Computer Science and Engineering Department
Rabindranath Tagore University Bhopal
yadavpritaj@gmail.com

ABSTRACT

Data visualization refers to studies that schematically represent abstracted data visually while simultaneously including their own features and factors. Data visualization's primary goal is to use graphical tools to communicate information in a more understandable and efficient manner. In order to come up with good results, data analysis is required. Decisions can be made using the final result. In order to help the human brain better understand and inspect the data, which is to comprehend the outcomes of their processing, data visualization entails turning the data into a visual context. The fastest human cognitive pathway is the visual canal. Because of this, more individuals prefer to view data than hear or read it. Fast data processing, analysis, and display are typically priorities in everyday life in a time of overproduction of digital data. Visual data analysis and processing tools are constantly updated as a result of constant contact between managers, users and developers on the one hand, and on the other, on the other hand. The essential points and the possible effects of data visualization tools and approaches in diverse fields are summarized in the chapter's conclusion. The purpose of this chapter is to present a thorough review of data visualization methods and techniques and to emphasize their significance in many disciplines for efficient data analysis and communication.

I. INTRODUCTION

A method of preparing and showing data known as analysis of data visualization (ADV) is used to make data more understandable for analysts, managers, and anybody else who works with managing and using large data in their day-to-day operations. Data visualization technologies, which make it possible to analyze as much meaningful data as possible and assist users in making decisions, are of special importance to big data analysts [1]. Data analysis is a method for analyzing and condensing data to reveal insights and spot patterns or trends. It is frequently the first stage of data analysis and helps determine the best statistical models by helping to comprehend the structure of the data, find outliers and anomalies, and understand how the data is organized. In order to comprehend the qualities, properties, and underlying structure of the data, this method of data analysis entails studying and summarizing the data. Data analysts can get insights into the data and spot patterns, trends, and relations that might not be immediately obvious according to this flexible and iterative method. Data Analysis utilizes a number of methods, such as statistical summaries, visualization, and hypothesis testing, which help analysts in better comprehending the data. In order to produce hypotheses that can be tested with additional research, Data Analysis aims to identify intriguing aspects of the data. Data Analysis is particularly useful for assisting analysts in spotting potential issues with the data, such as missing or wrong values, outliers, or discrepancies. Analysts may assure the accuracy and dependability of their findings by spotting these problems early on and taking action to resolve them.

Data visualization's primary goal is to use visuals to convey information in a more effective and clear way. When information is represented in visual forms, the human brain is better equipped to spot links and patterns and understand them. We can analyze data and spot patterns with graphs in a way that is not possible with any other method. Various graphic formats, such as bar graphs, pie charts, tables, and diagrams, can be used to visualize data. Technology has advanced quickly because of the usage of IT tools, but data analysis has also benefited from the expanding use of visualization, or making information more palatable in a visual manner [2]. Because the ordinary person's brain memorizes visual representations and information that it will get from the world through the visual sense fast, visualization itself is founded on the quick perception of visual forms by people. Finding methods and

techniques to improve daily human existence is a growing area of scientific interest. The rule that guides competition amongst the data visualization tools on the market is that they must be "closer" to users, i.e., more user-friendly. In order to better comprehend and communicate complicated data, data visualization is an effective tool. It entails displaying data in a graphical or pictorial format to make it simpler to comprehend and interpret. Data visualization is becoming more crucial than ever due to the growing availability of data in a variety of fields, including business, social sciences, humanities, sports, environmental sciences, and healthcare. This chapter offers a thorough introduction of data visualization tools, techniques, and their uses in many fields. In order to give readers an understanding of the various kinds of data visualization tools and techniques available, this chapter tries to emphasize the value of data visualization in clearly communicating and analyzing data. Whether you are a learner or a seasoned professional, this will be a useful tool for enhancing your comprehension of data visualization and its applications.

II. RELATED WORK

Many researchers and analysts regularly employ and, at the same time, update with new features the capabilities of the top tools for visual data analysis. We'll go over the most popular strategies, procedures, and equipment employed by big data researchers and analysts in order to summaries some of their findings. The emergence of potent interactive visual data analysis tools, made available through the various user-friendly programmes on the market, has advanced development in this field. The exploration and summarization of data to find patterns and relationships is a crucial first stage in the data analysis process. Data analysis will help data analysts in developing a deeper knowledge of the data, spotting potential issues or mistakes, and producing hypotheses that might direct additional study.

They emphasize the value of employing graphs and tables for graphical visualization while analyzing huge data. Their particular research is on how religious ceremonies are organized, including the services that are needed and provided. There is a steady flow of believers seeking detailed information on the facts the Hajj organizers required for their planning. Based on more in-depth service visualization, the article demonstrates how to use big data visualization efficiently to identify areas that require improvement [3]. The study makes it clear that typical visual representations on a map do not offer information that concentrates on enhancing services by acting appropriately at the appropriate moment. To identify issues and direct decision-makers' attention towards accurate assessment in order to improve services, advanced visual data analysis techniques are required. An effective method for examining the elements impacting the general public and the media was investigated by Kennedy et al [4]. The research suggests, among other things, linking the paradigms for media and communication studies and HCI (human-computer interface) to raise people's awareness of the usefulness of data visualization as a tool. The paper's conclusion discusses the need for better public understanding and interpretation of visual data, i.e., for various societal groups. Large-scale data visualization presents a number of technical difficulties, which Jagadish et al [5]. explored. It also discusses the difficulties presented by big data visualization while establishing a number of key terms related to big data and their visualization. Big data is also explained as a given in today's way of life, in business, and in their use by the general public. The researcher has made basic progress towards laying the groundwork for spatial analysis for Geographic Information Systems (GIS), for future data processing and more efficient data processing and decision-making, mainly in the health sector. Briefly stated, the study includes geographically mapped systems that support general public health in the United States, and as such, they can be regarded as the forerunner of current data visualization. Regarding big data, its processing tools, analysis, and decision-making procedures, reaches vast amounts of structured, unstructured, and semi-structured data, or, to put it another way, heterogeneous data. Their study focuses on how big data affects the healthcare industry and uses tools like Hadoop to explore the theoretical framework of data analysis. The utilization of Hadoop and MapReduce is the foundation of the research. Data visualization is seen as a crucial component of big data analysis in the research of Fehmi Skender. The study looks at efficient ways to visualize massive data, concentrating on interactive processes. The analyst is able to request several visualizations during an interactive visualization session, and each visualization builds on the one before it. The work focuses on merging distributed data processing systems that can efficiently handle massive data with a visualization system, even efficiently interactive visualization for lesser amounts of data. Additionally, the emphasis is on looking for solutions to reduce the length of time the visualization takes to complete. All of this is viewed as a conclusion that also serves as a potential response to upcoming needs and a starting point for analysts to get new experiences that would boost their productivity [6].

Overall, research emphasizes the value of data analysis and visualization and stresses the necessity of adaptable and iterative ways to investigating and summarizing data. Analysts can better comprehend the data and create more

precise and trustworthy models and insights by establishing hypotheses and examining the data in a systematic and iterative manner.

III. OVERVIEW OF DATA ANALYSIS METHODS

Common methods that are used for data analysis include:

A. Data Exploration:

The process of analyzing and comprehending data in order to gain knowledge and spot trends or links is known as data exploration. Data analysis entails using a variety of methods and tools, such as statistical analysis, visualization, and summary.

B. Data Cleaning:

Finding and correcting mistakes, inconsistencies, and inaccuracies in a dataset is known as data cleaning. It entails a number of duties, including processing missing data, updating data format, identifying and deleting outliers, dealing with duplicates, and addressing anomalies in data values. We applied methods for imputing missing data, including mean imputation, duplicate removal, format standardization, and data value correction.

C. Data Modeling:

The term "modeling" describes the process of turning raw data into a mathematical model that can be used to predict the future or categories brand-new data points. In machine learning, the aim of data modeling is to produce a model that accurately reflects the relationships and patterns in the data and can be applied to forecast the behavior of new, unforeseen data. Utilizing a linear regression model, we conducted statistical analysis. It entails a number of steps, such as:

- Data preparation: Data cleaning, transformation, and scaling are all examples of data preparation.
- Feature selection: choosing the variables or characteristics that are most pertinent to the model.
- Model selection: It refers to picking the best model or algorithm for the given data and issue.
- Model training: entails setting the model's parameters and training the model using the data.
- Model evaluation: involves verifying the model's performance using a different set of data and making any improvements.

The model can be used to generate predictions on current, unknown data after being trained and assessed.

D. Data Visualization:

The ability to explore the relationship between variables and spot patterns in the data makes data visualization an essential component of linear regression analysis. Scattered plots, regression lines, residual plots, and diagnostic plots were among the visualizations we employed. These can all be used in linear regression research.

E. Statistical Summaries:

Finding patterns and relationships in the data can be aided by computing summary statistics like mean, median, variance, and correlation coefficients.

F. Hypothesis Testing:

Analysts can assess whether patterns or relationships they detect are statistically significant or just the result of chance by testing hypotheses about the data.

IV. OVERVIEW OF DATA VISUALIZATION METHODS

To show data in a graphical or pictorial format that is easier to grasp and analyze, data visualization techniques are utilized. We also have a few standard procedures for data visualization, such as:

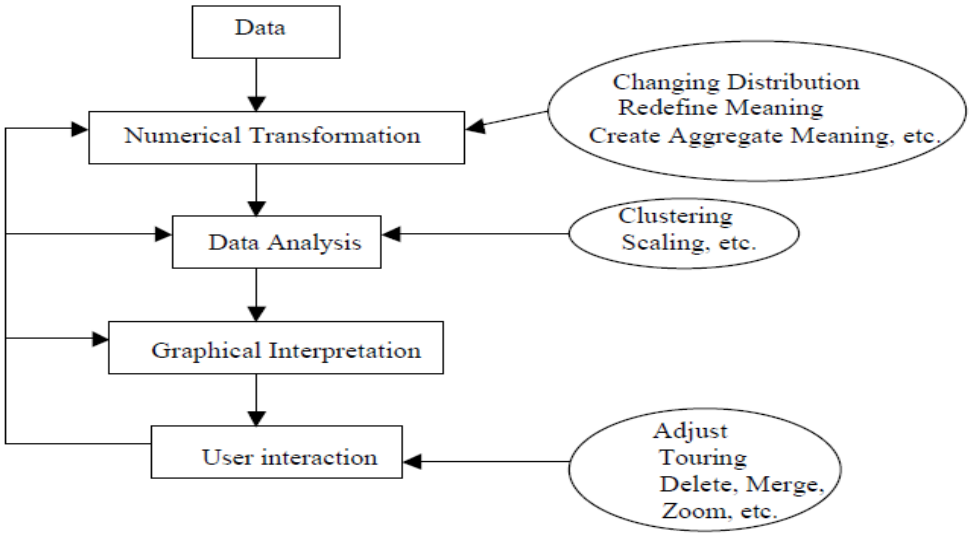

Figure1: Steps in data visualization

Data visualization approaches come in a variety of forms, such as:

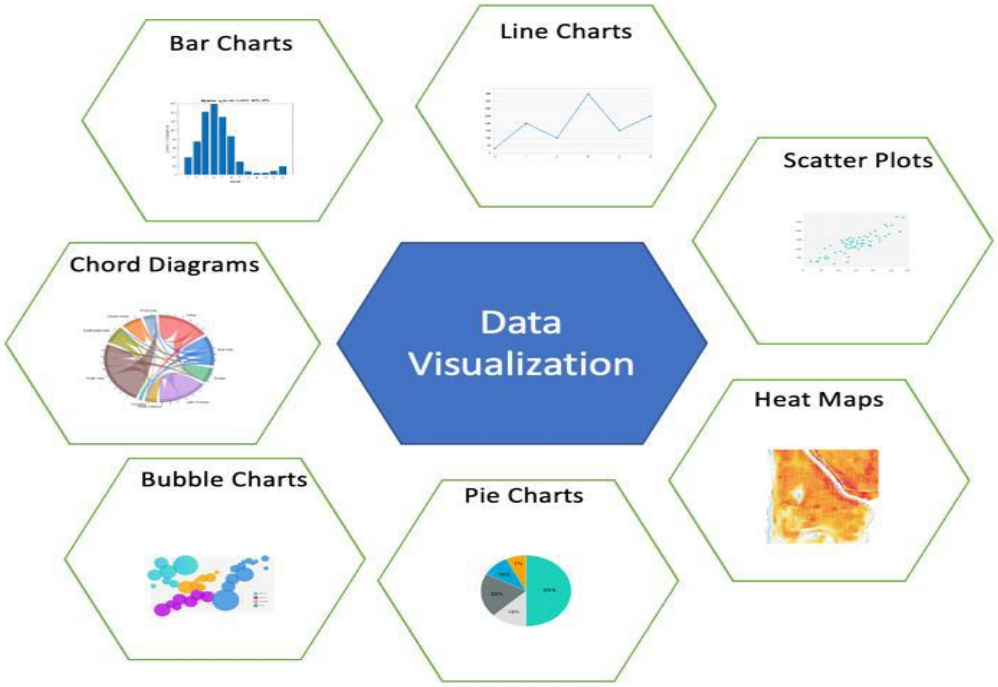

Figure 2: Data Visualization Approaches

Different data types can be represented using a variety of data visualizations, including bar charts, line graphs, scatter plots, heat maps, and network diagrams.

- Bar graphs, in which the length of each rectangular bar denotes the magnitude of a specific data point, are helpful for comparing data across categories.
- With a succession of points connected by lines, each representing a data point for a certain period, line graphs are used to visualize trends over time.

- On the other hand, scatter plots are used to show the relationship between two variables. Two values for two variables are represented by a succession of dots.
- The distribution of values in a matrix or table can be visualized using heat maps. The magnitude of a certain data point is represented by the color of each cell in a colored grid. Network diagrams, which have nodes for entities and edges for relationships, are useful for visualizing interactions between entities. The size and color of the nodes and edges can be used to represent different types of data.

V. OVERVIEW OF DATA VISUALIZATION TOOLS

Data scientists and the scientific community that employs this knowledge in practice are very interested in the constantly expanding subject of data visualization since it enables them to make more in-depth, quicker, and more effective observations that will enable data penetration. This will support them in making decisions that are more effective, quicker, and more helpful for their organizations. Therefore, it's crucial to present the data in a creative way, using simple tools like colors, elements, and dimensions, as well as analyses that have an impact on the data's representativeness. Numerous visualization technologies have been created and are employed because data visualization, particularly with regard to large data, necessitates not only an understanding of design and data but also a foundational understanding of statistics.

A. Tableau Public:

With Tableau Public, any type of visualization can be carried out quickly and easily without the need for coding knowledge. You must purchase a license and use the full paid version of the software in order to access all of its features. The options are wider in the commercial edition of the Tableau application, and it is also possible to mix several data types that are present in Excel or PDF formats. In summary, data may be visualized using Tableau Public by making simple tables, lists, maps, and a variety of other tools that are particularly good for visualizing massive data as well as for interactive data analysis between the data and the visualizations itself [7]. In the Tableau Alive version, the design makes it simple to connect to Hive (Big Data Business Intelligence) resources, enabling online data analysis without the need for data transfer (when data exceeds an order of five bytes in size). The architecture of Tableau Live safeguards investments in data warehouses and large data. Tableau can also display massive amounts of data by making a direct connection to the running database in memory. With incredible speed, capacity, and streaming capabilities, Tableau makes it possible to analyze hundreds of millions of rows of data.

B. Yellowfin Bi:

A platform for analytics that focuses on spreadsheets and visual analysis is called Yellowfin BI. It includes a large library of pre-made spreadsheets as part of its functionality as a contemporary tool for data visualization. The structure of the product enables users to perform more effectively. As a result of automated insights and machine learning algorithms, features of the Yellowfin BI tool [8] are based on machine learning, delivering on-demand visualization and highly clear and presentable visualizations. Another benefit of using this application is the ability to completely manage the visualization of data points while also obtaining focused results by choosing or eliminating specific data points in a spreadsheet.

C. Sap Analytics Cloud:

A tool for data visualization in business research, Sap Analytics Cloud is primarily made for planning businesses and examining their errors and decision-making. At any point throughout processing, different data may be used again. Sharing information in PDF format with other SAP Analytics Cloud users and external collaborators enables hands-on collaboration in SAP Analytics Cloud [9]. The SAP Cloud Analytics Story programme enables the use of a variety of expert, interactive, significant lists, and other items to appropriately show your data. Data may be made unique by using custom filters on spreadsheets and widgets, and bookmarks can be inserted into certain views for later research. The built-in calendar enables assignment, planning, and status tracking as well as passing reminders for the same.

D. Oracle Analytics Cloud:

With the capabilities of the machine learning tool [10], oracle analytics cloud is a data visualization solution for joint reporting with analytics for organizations of all data sizes, enabling the discovery of crucial information through simple data visualization. One of the main benefits of oracle analytics cloud is the processing of data in natural language, i.e., there is no requirement for good coding knowledge to find and analyze user-interesting data. The oracle analytics cloud tool has the following features: It can automatically build visualizations by discovering techniques based on machine learning in big data analysis. The tool's user-friendly interface allows identical data analysis using different data visualization approaches and processes.

E. DOMO:

DOMO is a Business Intelligence (BI) platform that runs in the cloud and has thousands of built-in visualizations, including over 1,500 list kinds and over 7,000 maps. By notifying and posing inquiries in plain language, machine learning enables automated data discovery. Deeply post-data genesis, the DOMO tool may access, filter, sort, and group data. You can use the tool to change the colors, graphics, and text in your layout.

The DOMO tool's features include the ability to combine separate perspectives, such as merging sales, return on investment, performance measurement, and any other key performance indicators that may be selected in one interface. The individual visualizations from the table can be divided with the use of the Pop-Up menus. The tool gives users the option to alter the visual display's appearance by choosing alternative list kinds, data series, and filters. This allows for interactive visualization enhancement.

F. POWER BI:

Power BI is a technology that enables the transformation of disparate data sources into standardized, visible, and interactive insights. It is a collaborative software service, application, and connection. The information may come from hybrid systems, Excel, a collection of cloud data, or a variety of other sources. Connecting to current data sources, visualizing the results, and sharing them among people or in the cloud are all made simple by Power BI. The primary benefit of using Power BI as a data visualization tool is that it is readily available and reasonably priced. Power BI Desktop is cost-free. Downloading, setting up, and using it are simple. The use of Power BI allows for simple bespoke visualizations in spreadsheets and reports as well as interactive visualization. Key performance indicator visualization, maps, charts, graphics, R image scripts, dashboards, and other tools are among the features. The flexibility to transfer data from a variety of sources and link data to XML and JSON formats is a significant benefit of utilizing Power BI as a data analysis tool.

VI. CONCLUSION

In conclusion, data visualization is an effective method and approaches that enable analysis and comprehension of complicated data, the identification of patterns and trends, and the dissemination of insights to a larger audience. A clear and visually appealing presentation of the data is necessary for effective data visualization, as is the selection of the best tool and technique for the data and message being communicated. Data visualization supports well-informed decision-making, enhances results, and advances knowledge and comprehension in a variety of sectors. It is impossible to emphasize how vital data visualization is, and the future is likely to see it play a bigger and bigger role across a variety of fields. Along with the aforementioned advantages, data visualization also has the ability to democratize knowledge and make it available to a larger audience. The number of data visualization tools is expanding quickly, and they are continually enhancing and expanding their analytical skills, particularly in the area of Big data. Given that 90% of the information presented in the brain is visual and that the human eye is drawn to visual representations, colors, and patterns, it is clear that data visualization is the most advantageous technique for business and, generally, for daily needs for visual data analysis in any area of life. The trend of using visualization tools more frequently is growing as a result, and they are now more important than ever.

REFERENCES

[1]. Stephen R. Midway, “Principles of Effective Data Visualization”, Pattern Perspective, ScienceDirect, Volume 1, Issue 9, pp. 1-7, 11 December 2020.

[2]. Pradnya Khot, Aniket Milkhe and Gayatree Sorte," Data Visualization- Techniques and its applications", International Research Journal of Engineering and Technology, Volume 08, Issue 07, pp. 2974-2978, 2021.

[3]. Abdulkhaliq Alharthi, Vlad Krotov and Michael Bowman, "Addressing barriers to big data", ScienceDirect, Volume 60, Issue 3, PP. 285-292, May–June 2017.

[4]. Kennedy, H., Weber, W. and Engebretsen, "Data visualization and transparency in the news", In: Engebretsen, M. and Kennedy, H., (eds.) Data Visualization in Society. Amsterdam University Press, pp. 169-185, 2020.

[5]. H.V. Jagadish, "Big Data and Science: Myths and Reality", Big Data Research, ScienceDirect, Volume 2, Issue 2, PP. 49-52, June 2015.

[6]. Skender, F., Manevska, V. (2022). Data Visualization Tools - Preview and Comparison. Journal of Emerging Computer Technologies, 2(1), Volume:2, No:1, Pages: 30-35, June 2022.

[7]. Steven Batt, Tara Grealis, Oskar Harmon and Paul Tomolonis, "Learning Tableau: A data visualization tool", The journal of economic education, vol. 51, nos. 3–4, pp. 317–328, 2020.

[8]. Mihaela Filofteia Tutunea,"Business Intelligence Solutions for Mobile Devices - an Overview",ScienceDirect,22nd International Economic Conference – IECS 2015 ,Economic Prospects in the Context of Growing Global and Regional Interdependencies, pp.160-169, IECS 2015.

[9]. Kjetil Hansen, Moutaz Haddara and Marius Langseth, "Exploring Multi-Criteria Decision-Making Methods in ERP Selection", ScienceDirect, Procedia Computer Science 219, pp. 879–888, 2023.

[10]. https://docs.oracle.com, "https://docs.oracle.com/en/cloud/paas/analytics-cloud/visualize-data.html," 6 2022.