Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777730

A REVIEW ON DATA VISUALIZATION FOR EXPLORATORY DATA ANALYSIS

Hari Prasad Josyula[1]*; Kaushikkumar Patel[2]; Amit Bhanushali[3]; Sunil Rajaram Landge[4]; Shobhit Mittal[5]

[1*]Hari Prasad Josyula, MBA, CSPO, Senior Product Manager – Fin Tech, Princeton, New Jersey
Email: jhprasad@outlook.com

[2]Kaushikkumar Patel, Independent researcher, Director Data Development, White Plains, NY

[3]Amit Bhanushali, Independent researcher, West Virginia University, West Virginia

[4]Sunil Rajaram Landge, Independent researcher, Technical Consultant, New York

[5]Shobhit Mittal, Independent researcher, Financial Management and Business Ops Expert, Cedar Park, Texas

***Corresponding Author**

*Hari Prasad Josyula, MBA, CSPO, Senior Product Manager – Fin Tech, Princeton, New Jersey, Email: jhprasad@outlook.com

Abstract

This study is aims to show the relevance of data visualization in data analysis. The presentation has focused on discussing and demonstrating how different data visualization techniques can be considered analyzing complex dataset. The demonstration is based on human resource dataset about employees. The data visualization performed in this study is done through Microsoft Power BI and presentation is done through charts and tables.

Keywords: Data visualization, analysis, chart, graph, exploratory data analysis

1.0 Introduction

Data visualization is a process in data analysis that entails transformation of data into visual form for ease of presentation and interpretation (Khot et al., 2021). The need for data visualization as a key process in data analytic process is underscored by the huge amount of data that various entities including government and business deal with on daily basis (Qin et al., 2020). Essentially, the various types of data visualization techniques such as tables, graphs, pie charts, patterns, network diagrams, tree maps, and word clouds empower data analysts and other users of data to arrive at accurate interpretation of a given set of data (Agbehadji & Yang, 2020; Ledford et al., 2019). The fact that data visualization enables data analysts and researchers to use graphical illustrations to represent relationships explains why the technique is considered one of the key building blocks in Exploratory Data Analysis (Ledford et al., 2019). In essence, the growing utilization of data visualization across different fields and by different entities is informed by the need to present data in a format that is appealing and can be interpreted easily by different users of a given dataset (Parmentier-Cajaiba & Cajaiba-Santana, 2020).

1.1 Research Problem

The number of studies exploring the role and utilization of data visualization in data analysis including the various techniques that can be used to visualize and present data in exploratory data analysis (EDA) process have been on the rise (Agbehadji & Yang, 2020). Despite data visualization emerging as an important process in EDA, the number of data visualization techniques covered in majority of these studies is limited. Preliminary review of existing studies indicate majority of existing studies rely majorly on graphical illustrations such as bar graphs, line graphs, scatter diagrams and pie charts techniques of data visualization (Khot et al., 2021). With visualization developing into a critical component of exploratory data analysis (Rawat et al., 2021), this study seeks to provide a comprehensive understanding data visualization techniques. In other words, the study attempts explain how each of the different data visualization techniques including tables, graphs, patterns, network diagrams, tree maps and data clouds can be employed to represent complex data (Gandhi & Pruthi, 2020).

1.2 Research Objective

The principal objective of this study is explore and provide a comprehensive understanding of the role data visualization in exploratory data analysis (EDA). In particular, the study seeks to explain the use of various types of data visualization techniques in data analysis process.

2.0 Literature Review

2.1 Overview of Data Visualization in EDA

Data visualization is one of the key processes in exploratory data analysis (EDA). As key component of EDA, data visualization encompasses presentation of a given set of data in a visual or graphical format (Rawat et al., 2021). According to Qin et al., (2020) data visualization process entails utilization of visual images or graphics to generate particular insight from a dataset under investigation. Data visualization is considered a key process in exploratory data analysis because it allows individuals to make informed decisions from a given set of data quickly and effortlessly (Agbehadji & Yang, 2020). Visual data is also considered critical to data analysis process because it enables researchers to gain comprehensive understanding of specific attributes present in a given dataset (Gandhi & Pruthi, 2020). The increasing relevance of data visualization in data analysis is also underscored by the fact data presented in visuals or graphics makes it easier for data users to identify inconsistences, patterns and relationships in a given dataset (Ledford et al., 2019).

The increasing popularity of visualization in data analysis is also underscored by the fact data visuals or graphics make it easier to identify patterns and hence relationships in a given dataset (Parmentier-Cajaiba & Cajaiba-Santana, 2020). In particular, Khot et al., (2021) argues data presented in visual or graphic form such as scatter plots or line graphs is quite effective in establishing trends and relationships in a given dataset. Also, data visuals including data summaries, scatter diagrams, cloud maps, and graphs are quite effective in presentation of complex and huge amount of data (Agbehadji & Yang, 2020). The fact that data visuals are quite effective in establishing relationships explains why they are often utilized in hypothesis testing in research studies. Visuals are also considered critical to the data analysis process because they can be used to identify inconsistent, inaccurate and misleading data from a particular set of data.

2.2 Data Visualization Techniques

Summary Tables

Summary tables is one of the most commonly used data visualization techniques. Summary tables entail presentation of key data statistics such as mean, median, range and standard deviation in a table (Gandhi & Pruthi, 2020). Rows and columns are often used to represent specific attributes of data being investigated. Table summaries are often used in the classroom environment to present student performance in difference subjects.

Graphs

Graphical illustrations are often used to show data summaries, show trends and portray relationships between different set of variables in a dataset. Typically, a graph has two axis, the y and the x axis. In vertical graphs, the y –axis is used to capture dependent attributes while the x-axis is used to capture independent attributes (Gandhi & Pruthi, 2020). On the other hand, horizontal graphs has the y-axis representing the independent attribute and the x-axis representing the dependent variables. The length of the bars, whether in vertical or horizontal graphs are often proportional to the attributes being measured or illustrated.

Pie charts

Pie chart is another popular data visualization technique. Pie-charts are useful where the aim is to provide a snapshot of the size of different data categories at particular period (Gandhi & Pruthi, 2020). In a pie-chart, the size of different attributes is often expressed in degrees. Just like bar graphs, the size of the pie are often proportional to the size of the attribute being compared and presented in a pie chart.

Line Graphs

The design and architecture of line graphs is similar to that of bar graphs. However, instead of bars, line graphs utilize lines to indicate the size of attributes included in the presentation (Khot et al., 2021). These line can also be used to show trends and patterns. Because of this line graphs can easily be used to show relationship between two sets of variable.

Scatter plots

Just like bar graphs and line graphs, scatter plots have two key axis, the horizontal or x- axis and vertical or y-axis. Use of two axis implies that scatter plots can equally be used to depict the relationship between the two attributes being compared. One advantage of scatter plots is that they can equally be used to indicate dispersion around the mean.

Tree Maps

This data presentation technique entails showing data in a hierarchy format (Gandhi & Pruthi, 2020). Tree maps diagrams are often used to show different stages in a process. In other words, tree maps are used to provide snapshot representation of complex processes.

Heap maps

The technique entails the use of different colors and densities to express particular attributes (Gandhi & Pruthi, 2020). The technique is often used by political strategies and analysis to categorize the campaign trail into favorable, contested or hostile. The technique is also commonly used in weather forecasts programs. Within the health field, the technique is often used to illustrate areas infested by diseases during epidemics. A scale indicating what the different colors represent is often used to help users make informed decisions of heap maps.

Word Clouds

This technique of data visualization entails use of key words to assess various sentiments. Word cloud mining entails using software programs to identify key word often used by social media and internet users to express specific sentiments. The significance of each word in a word cloud is expressed by font size and density used. Colour may also be used to illustrate significance of key words.

2.3 Empirical Review

Khot et al., (2021) investigated data visualization techniques and areas where they are commonly utilized. The study expressed data visualization as a technique of data presentation that entails the use of visuals to illustrate specific attributes. The specific data visualization techniques investigated in the study include tree diagrams or flow charts, line graphs, bar chats scatter plots and maps. While visual diagrams can be used to visualize large amount of data, the effectiveness of the various techniques is affected by audience, content and context of data being visualized. Nevertheless, the researchers conclude that at a time Big Data is the norm, visualization techniques help users draw useful insight from specific datasets.

Agbehadji and Yang, (2020) sought to investigate the effectiveness of various data visualization techniques in presentation of data. In particular, the researchers to determine the most effective methods for presenting animal behavior. Findings from the study show traditional methods of data presentation such as line graphs and bar graphs require a lot of time during construction and interpretation process. The researchers advocate for use of data mining algorithms as being quite effective as they do not involve computations. Algorithms are also better suited to capture qualitative data as opposed to traditional methods that rely heavily on quantitative data.

Gandhi & Pruthi, (2020) present data visualization as dynamic and highly effective method of presenting data. The study details attributes of various data visualization techniques and how each can be used to represent data. Some of the data visualization techniques covered in the study include cloud words, tree maps, scatter diagrams and graphs. According to the text data visualization is highly effective in presentation of enterprise specific data. Gandhi and Pruthi (2020) conclude that data visualization techniques are quite effective as users can draw conclusions within short period.

3.0 Methodology

3.1 Data source

The data used for demonstration of data visualization was derived from a human resource human resource database. The data was selected for the purpose of demonstration of the way data can be integrated and combined to summarize large chunks of information in a way that makes meaning to the reader. The dataset in consideration contains employee data, human resource department analytics data and data on promotion and next retrenchment of employees. The data is considered useful because it allows review of different techniques to be considered for visualization purpose. The data files obtained from the company database are in the comma separated format. The database constitutes of 1470 rows about employees.

3.2 Proposed Data Analysis and Visualization

In the case of this study, data visualization was completed through the use of Microsoft Power BI. It is an analytical software from Microsoft that allows organization of data and visualization through a number of techniques. The tool is preferred because it easier to use and also has the capacity to handle thousands of records. It is one of the data visualizations tools that are considered easy for use because it is mainly allows user to drag and drop. The data visualization process started by loading data into the Power BI, checked for errors and then actual visualization completed. Transformation process (cleaning of data) involved the use of Tab delimiter in order to organize data into relevant columns for further analysis. The visualization performed on the dataset focused on giving summary of information contained in the database, specific details for employees and actions to be considered for employees, that is either promotion or retrenchment. The presentation of data visualization is done through a combination of techniques including graphs, tables, and charts.

4.0 Analysis, Results and Discussion

The visualization performed in the dataset focused on three main aspects of the data, that is, overview summary of the data, details of the data and action for consideration based on the data. It is important to note that the discussion of the visualizations focused on demonstrating the aspects of visualization rather than getting into detailed discussion of the human resource aspects of the data presented. This section discusses the different visualizations performed and the way information is presented for the different purposes.

4.1 Dataset overview summary

Figure 1 is a snapshot taken from the visualization performed in eh Power BI to have an overview of data visualization. In the summary overview f summary presentation, it is evident that various techniques of visualizations have been used. Based on the different visualizations presented, it can be argued that there is adequate organization of information, allowing the audience have understanding of the whole data at a single glance. For instance, it is evident that that stacked column and bar charts and donut charts have been used organize number of employees based on the years of service, number of employees based on the job levels as well as categorization of employees based on their location distance from the office. Notably, the visualization include the use of shapes which have been customized to visualize specific information such as such as gender. The visualization on Figure 1 demonstrates the first aspect of data visualization techniques that can be achieved through the different visualization techniques.

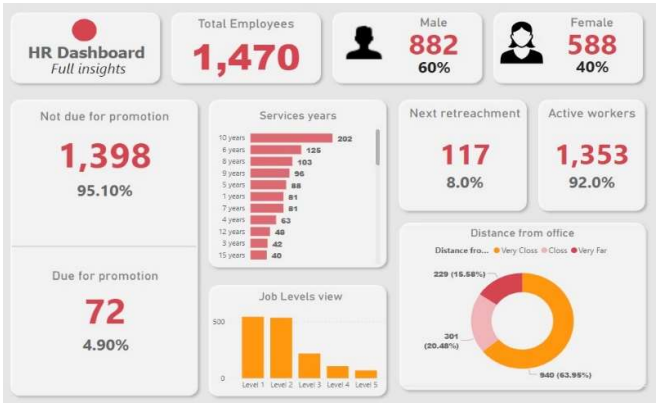

Figure 1: Summary of dataset

4.2 Details in Data Visualization

The second aspects of data visualization demonstrated in the analysis is the level of interaction that the audience can have with data when using different aspects of data visualization. It is important to note that visualization is more than creating summaries. In essence, data visualization is also about the level of interactivity that users can perform on the visualizations to realize how the data changes. Figure 2 and 3 represent similar kind of presentation with Figure 3 being filtered to show variation in the rest of the visualization.

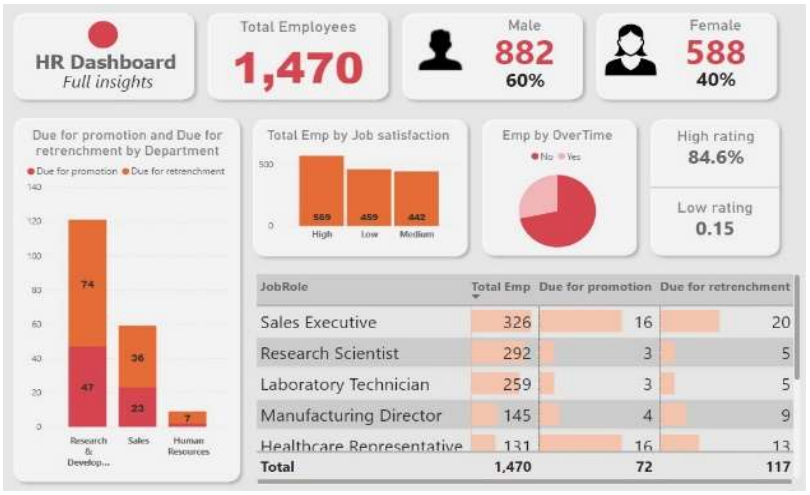

Figure 2: Detailed visualization analytics

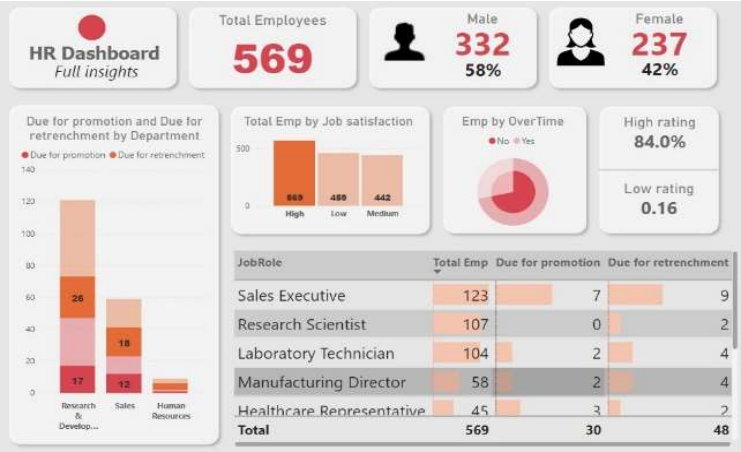

Figure 3: Demonstration of visualization interactions

Figure 2 and 3 is a demonstration of the levels of details about data that can be achieved through data visualizations. In Figure 2, it is evident that different visualization techniques can be combined to see different aspects of complex data that is being analysed. For instance, use the stacked column chart demonstrates the number of employee due for promotion and retrenchment based on their department. Further details are visualized through use of tables which compares the number of employees for retrenchment and promotion based on their job role. The emphasis is made through the inclusion of a bar chart on the table to help the audience easily compare the employees for promotion and retrenchment based on their role. In Figure 3, an interaction has been performed on the data by filtering presentation based on employees who are considered to be having job satisfaction. This interaction has an impact on all other graphs, allowing the audience to see all other aspects of data by considering employees who are highly satisfied by the job. For instance, only 332 males have high job satisfaction.

5.0 Conclusion

This study has focused on demonstrating the use of data visualization techniques and how such techniques contribute to understanding the data. It is important to note that findings of the study demonstrates that data visualization is relevant for organizing and presenting data in an easy to understand manners. Based on the findings of this study, it can be argued that data visualization techniques through the use of different tools such as Power Bi and Neural Networks are relevant in the increasingly complex business environment. In essence, the choice of data visualization and its suitability depends on the type of data and the aim that a researcher aims to achieve. The use data visualization techniques has demonstrated that various techniques create value by helping to simplify data in ways that are easy to understand.

References

[1]. Agbehadji, I.E. & Yang, H. (2020). Data Visualization Techniques and Algorithms. *Computer Science.*

[2]. Gandhi, P. & Pruthi, J. (2020). *Data Visualization Techniques: Traditional Data to Big Data.* Springer Nature Singapore Ltd.

[3]. Khot, P., Milkhe, A. & Sorte, G. (2021). Data Visualization- Techniques and its applications. International Research Journal of Engineering and Technology (IRJET), *vol. (8)7*

[4]. Ledford, J. R., Barton, E. E., Severini, K. E., Zimmerman, K. N., & Pokorski, E. A. (2019). Visual display of graphic data in single case design studies. *Education and Training in Autism and Developmental Disabilities*, *54*(4), 315-327.

[5]. Parmentier-Cajaiba, A., & Cajaiba-Santana, G. (2020). Visual maps for process research: Displaying the invisible. *M@ n@ gement*, *23*(4), 65-79.

[6]. Qin, X., Luo, Y., Tang, N., & Li, G. (2020). Making data visualization more efficient and effective: a survey. *The VLDB Journal*, *29*, 93-117.

[7]. Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights*, *1*(2), 100012.