Exploring Data Visualization in Pharmaceutical Evaluation

Abstract

The research paper delves into the realm of data visualization within the context of pharmaceutical evaluation, aiming to illuminate its significance in understanding and communicating complex drug performance metrics. By searching different papers from various domains, including business, social sciences, sports, healthcare, this paper underscores the versatile applications of data visualization tools and techniques. Through a comprehensive overview, the paper highlights the pivotal role of data visualization in enhancing data comprehension and analysis. Specifically focused on drug performance evaluation. In addition, by providing a comprehensive overview, it serves as a valuable resource for enhancing data communication and analysis in the realm of pharmaceutical evaluation.

Keywords

Data Visualization, Pharmaceutical evaluation, Drug performance

I. Introduction

In today's world, where data is everywhere and plays a big role in decision-making, the pharmaceutical industry is no different. It is important for companies making medicines to understand how well their drugs work. With so much information available, it is hard to make sense of it all. That's where data visualization comes in.

Additionally, data visualizations serve as powerful aids in uncovering intricate patterns and trends hidden within the dataset. By presenting complex data in a clear and accessible manner, researchers can derive significant insights into medication efficacy, usability, and patient satisfaction. Such insights are invaluable for guiding informed decision-making by healthcare professionals and stakeholders in the pharmaceutical industry.

In summary, our research aims to bridge the gap between raw data and actionable insights, thereby facilitating evidence-based practice and continuous improvement in pharmaceutical evaluation. Through meticulous data analysis and visualization, we aspire to contribute meaningfully to the advancement of healthcare quality and patient care outcomes.

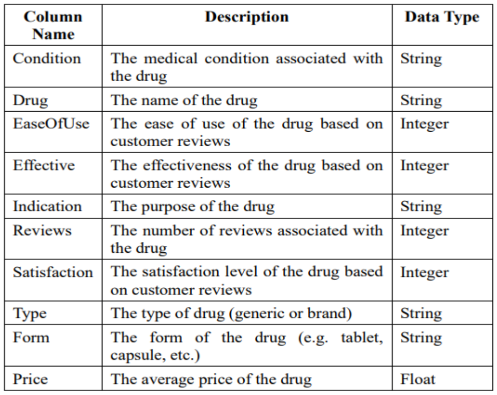
II. Overview of Dataset & Visualization

a. Data Collection

In the study, the data collection process involved sourcing preprocessed data from "*Kaggle*". Kaggle, a widely used platform for data science competitions and datasets, provided a refined dataset that had already undergone preprocessing. This dataset was meticulously prepared, ensuring that the data was clean, structured, and ready for analysis. By leveraging preprocessed data from Kaggle, researchers were able to focus their efforts directly on data analysis and interpretation, by passing the time-consuming task of data cleaning and preprocessing. This streamlined approach enabled researchers to efficiently explore and extract valuable insights from the dataset, contributing to the study's objectives and outcomes.

b. Dataset Overview

The dataset from Kaggle named "*Drug Performance Evaluation*" provides information on 37 common medical conditions and the drugs associated with them. Each drug entry includes details such as its name, type (generic or brand), form (tablet, capsule), average price, indications (purpose), customer reviews, ease of use and effectiveness, and satisfaction ratings. For instance, the "Condition" column specifies the medical condition targeted by each drug, while the "Price" column indicates its average cost. Understanding these attributes helps us evaluate the performance and usability of different drugs for various conditions. Here, the inclusion of consumer reviews serves as a key component of medication reception, offering valuable insights for healthcare decision-makers and consumers alike.



Column Name	Description	Data Type
Condition	The medical condition associated with the drug	String
Drug	The name of the drug	String
EaseOfUse	The ease of use of the drug based on customer reviews	Integer
Effective	The effectiveness of the drug based on customer reviews	Integer
Indication	The purpose of the drug	String
Reviews	The number of reviews associated with the drug	Integer
Satisfaction	The satisfaction level of the drug based on customer reviews	Integer
Type	The type of drug (generic or brand)	String
Form	The form of the drug (e.g. tablet, capsule, etc.)	String
Price	The average price of the drug	Float

Figure 2.1: Dataset Column Names & Description

c. Data Visualization Tool

The primary challenge had lied in finding the best ways to show important information from complicated datasets. Using the "R language" in the "R Studio" software, visualization techniques were essential for making complex data easier to understand. By turning data into simple pictures like charts and graphs, visualization helped us to see patterns and trends that might be hard to spot just by looking at numbers and strings in dataset. This method made it easier for researchers and decision-makers to understand the data better and helped to make informed choices, derive actionable insights from the data across different areas of study and industry.

III. Overview of Data Visualization Techniques

For the data visualization we used several kinds of methods.

a. Boxplot

A boxplot, also known as a box-and-whisker plot, is a statistical visualization tool that provides a concise summary of the distribution of a dataset. By visualizing boxplot plot in every column, we can learn about the dataset as well as its data.

In boxplot, numeric columns were used to check outliers. Here, the numeric column were extracted from the data set and stored into another variable called “numeric_colours”.

```
boxplot(numeric_columns,
    main = "Boxplot of Numeric Columns",
    xlab = "Variables",
    ylab = "Values")
```

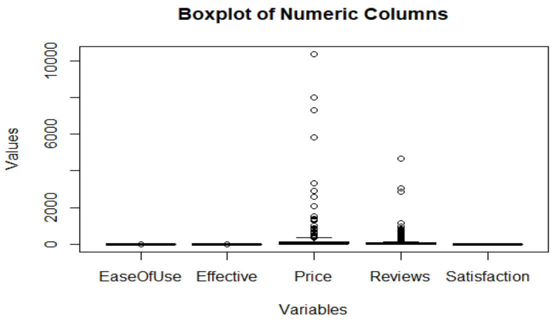


Figure 3.1: Box plot

In the dataset, the outliers are active in “Price” and “Reviews”.

b. Histogram

A histogram is a graphical representation of the distribution of numerical data. The histogram method of each numeric column of the dataset was applied. The par() function was used in R code to set or query graphical parameters. The mfrow parameter specified the number of rows and columns for the layout plots.

```
numeric_columns <- drug_data[, c("EaseOfUse",
"Effective", "Price", "Reviews", "Satisfaction")]

par(mfrow=c(2, 3))

for (col in colnames(numeric_columns))
{hist(numeric_columns[[col]], main = col, xlab = col)}
```

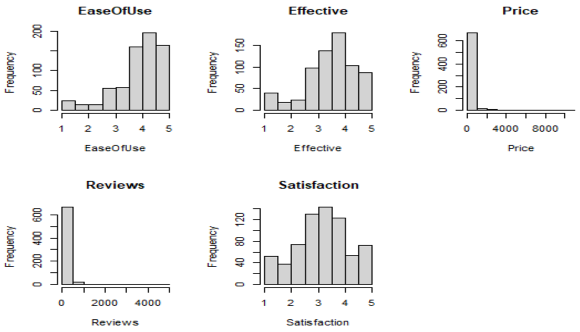


Figure 3.2: Histogram (Bar Chart)

In the dataset by this visualization, the frequency of each numeric column were detected.

c. Bar Plots

A bar plot is a graphical representation of categorical data in which rectangular bars are used to represent the frequencies or proportions of different categories. Bar plots were used to check the range of categorical data in each column which held categorical data.

```
categorical_columns <- drug_data[, c("Condition", "Drug", "Form", "Indication", "Type")]

par(mfrow=c(3, 2))

for (col in colnames(categorical_columns)) {

  category_counts <- table(categorical_columns[[col]])

  barplot(category_counts, main = col, xlab = "Category", ylab = "Frequency")

}
```

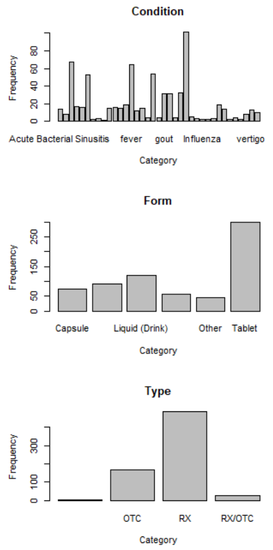


Figure 3.3: Bar plots

This visualization shows us the frequency of different attributes.

d. Scatter Plots

A scatter plot is a type of data visualization that displays the relationship between two numerical variables. It consists of a set of points, where each point represents the value of one variable plotted against the value of another variable.

The relation between numerical columns was checked. The pairs() function in R is used to create a scatterplot matrix, which displays scatter plots for each pair of variables in a dataset.

```
numeric_columns <- drug_data[, c("EaseOfUse", "Effective", "Price", "Reviews", "Satisfaction")]

pairs(numeric_columns)
```

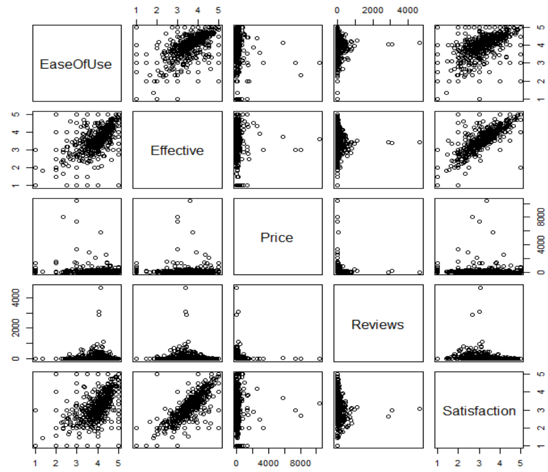


Figure 3.4: Scattered Plot

e. Line Plots

A line plot, also known as a line graph, is a type of data visualization that displays data points connected by straight lines. It is commonly used to show trends and changes over time or across different categories.

```
satisfaction_means <- tapply(drug_data$EaseOfUse, drug_data$Satisfaction, mean)

plot(as.numeric(names(satisfaction_means)), satisfaction_means, type = "o",

   xlab = "Satisfaction", ylab = "Mean EaseOfUse",

   main = "Mean EaseOfUse Across Satisfaction Levels")
```

Here, satisfactory_means was used to calculate the mean of EaseOfUse value for each level of satisfaction using tapply() function.

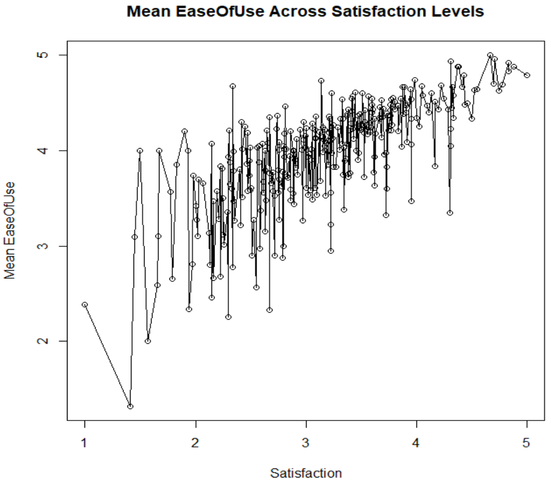


Figure 3.5: Line Plots

The relation between Mean Ease of use and satisfaction was visualized.

f. Heatmaps

A heatmap is used to make a graphical representation of the data. It is very useful in big data sets because of quickly identifies patterns or trends.

```
library(corrplot)

numeric_columns <- drug_data[, c("EaseOfUse", "Effective", "Price", "Reviews", "Satisfaction")]

correlation_matrix <- cor(numeric_columns)

corrplot(correlation_matrix, method = "color", type = "upper",

         addCoef.col = "black", tl.col = "black",

         title = "Correlation Matrix Heatmap")
```

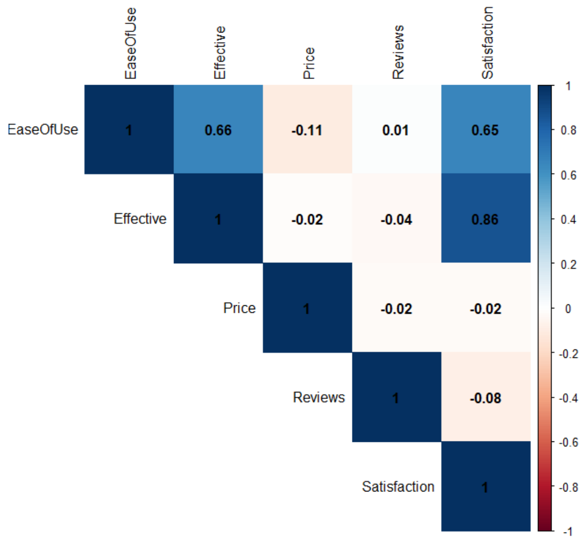


Figure 3.6: Heatmap

To use heatmap a library 'corrplot' was called. The correlation matrix used cor function to calculate the relation between the columns.

g. Violin Plots

This method is basically used in data visualization to visualize the distribution of numeric data and provides insights into the shape, spread, and central tendency of the data, like a box plot, while also displaying the probability density of the data at different values, like a kernel density plot.

```
library(ggplot2)

ggplot(drug_data, aes(x = Type, y = Satisfaction, fill = Type)) +

  geom_violin() +

  labs(title = "Satisfaction by Drug Type", x = "Type", y = "Satisfaction") +

  theme_minimal()
```

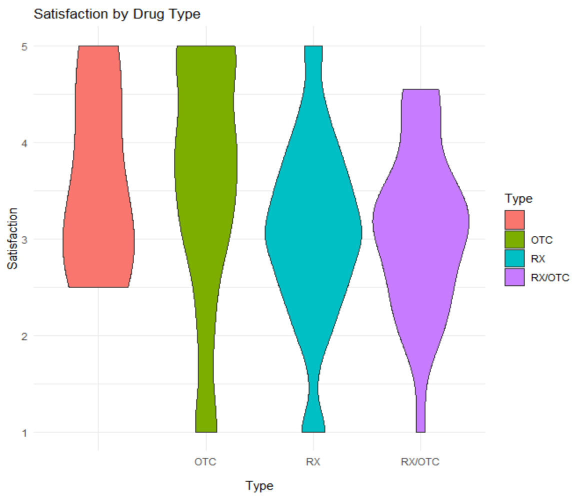


Figure 3.7: Violin Plots

h. Pie Charts

The pie chart method is most effective when we represent data that can be easily summed up to a meaningful total. A pie chart is a circular statistical graphic divided into slices to illustrate numerical proportions. Each slice represents a proportionate part of the whole, and the size of each slice is proportional to the quantity it represents.

It was used in categorical columns to explore the range of data.

```
categorical_columns <- drug_data[, c("Condition", "Drug", "Form", "Indication", "Type")]

par(mfrow=c(3, 2))  # Arrange plots in a 3x2 grid

for (col in colnames(categorical_columns)) {

  category_counts <- table(categorical_columns[[col]])

  pie(category_counts, main = col)}
```

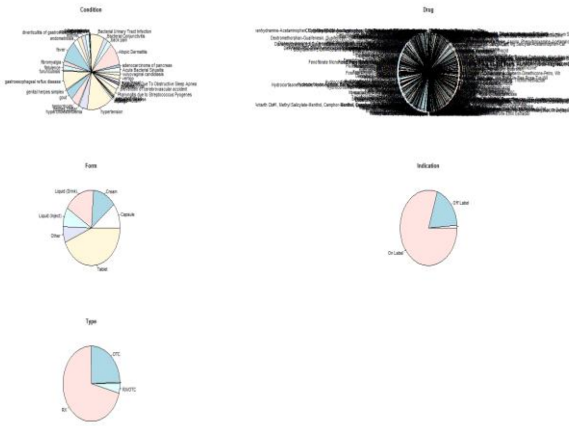


Figure 3.8: Pie charts

There is some data which holds a wide range of amounts.

i. Faceted Plots

Faceted plots, sometimes referred to as trellis plots or lattice plots, are a kind of data visualization in which several graphs are made using different data subsets. A distinct subset of the data is represented by each plot, or panel, making it simple to compare data across several groups or categories.

```
ggplot(drug_data, aes(x = EaseOfUse, y =
Satisfaction)) +

  geom_point() +

  facet_wrap(~ Condition) +

  labs(title = "Scatter Plot of EaseOfUse vs.
Satisfaction Faceted by Condition",

      x = "EaseOfUse", y = "Satisfaction")
```

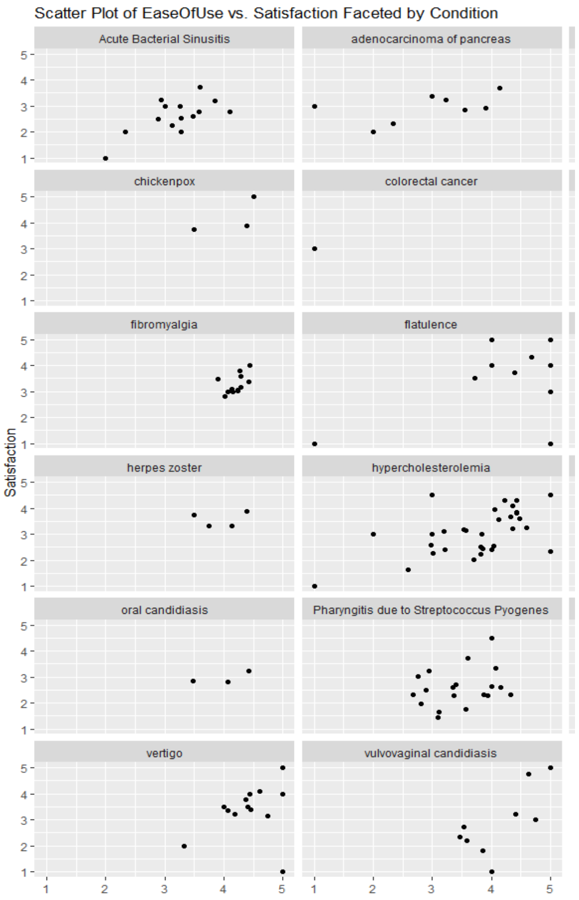


Figure 3.9: Faceted Plots

j. Bubble Plot

A bubble plot is a kind of data visualization in which data points are shown as circles or bubbles on a two-dimensional plot. Each bubble's size and position correspond to two and three variables, respectively.

```
ggplot(drug_data, aes(x = Price, y = Reviews, size =
Effective, color = Effective)) +

  geom_point(alpha = 0.5) +

  scale_size_continuous(range = c(1, 10)) +

  labs(title = "Bubble Chart of Price vs. Reviews by
Effectiveness", x = "Price", y = "Reviews") +

  theme_minimal()
```

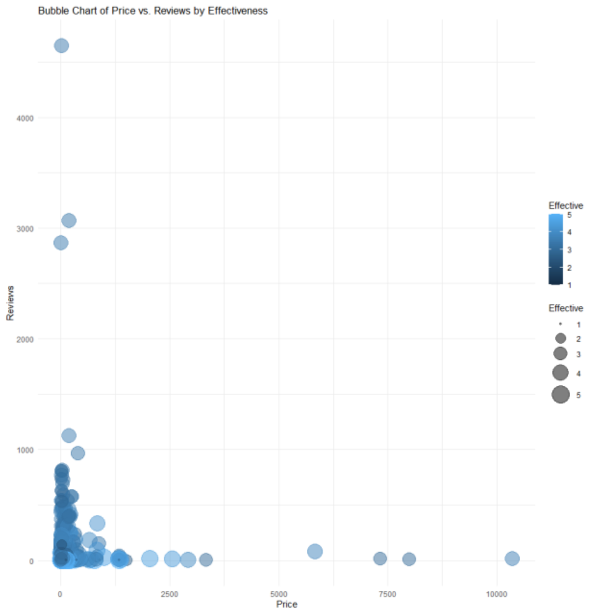


Figure 3.10: Bubble Plot

k. Geom_bar plot

A geom_bar plot, a popular plot type used to generate bar charts in R using the ggplot2 library.

```
library(dplyr)

library(ggplot2)

condition_form_count <- drug_data %>%

  group_by(Condition, Form) %>%

  summarise(Count = n(), .groups = 'drop') %>%

  arrange(desc(Count))

ggplot(condition_form_count, aes(x =
reorder(Condition, Count), y = Count, fill = Form)) +

  geom_bar(stat = "identity") +

  labs(title = "Distribution of Medication Forms by
Condition", x = "Condition", y = "Count") +

  theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

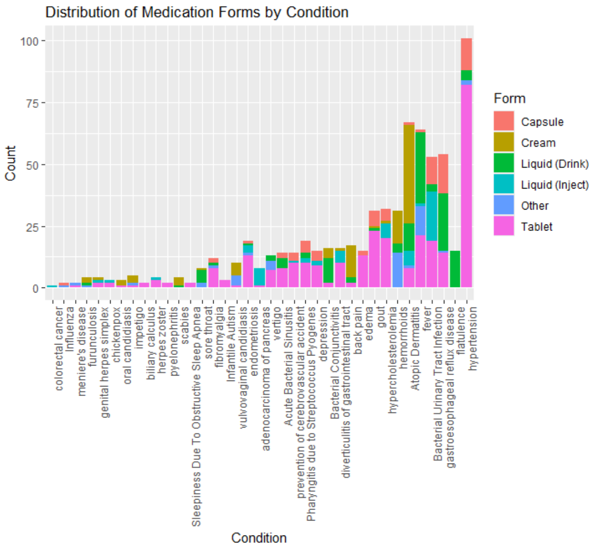


Figure 3.11: Geom_bar Plotss

IV. Challenges in Data Visualization

Challenges raised when dealing with categorical data, such as the "Condition" column indicating medical conditions associated with drugs, where the diversity of categories overwhelmed traditional visualization methods. The pie chart that represented the wide range of medical conditions resulted in a cluttered and difficult-to-interpret visualization due to the huge number of categories.

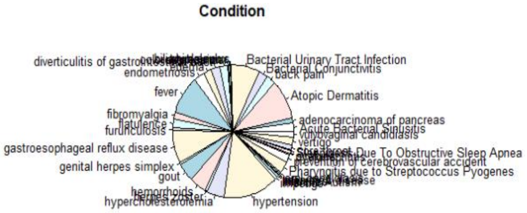


Figure 4.1: Pie-chart of "*Condition*" column

Wide variation in scale and distribution of numeric data, like line plots gave some complicated visualization selection and interpretation.

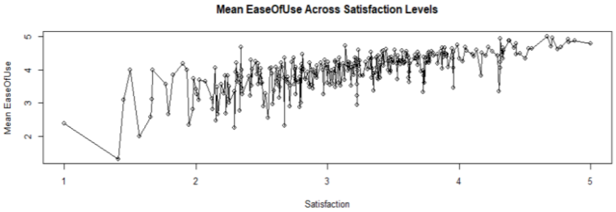


Figure 4.2: Line plot of "*MeanOfEase*" & "*Satisfaction*" column

Despite the challenges that arrived, this study provides valuable experience in handling data visualization and enhancing understanding. By navigating through complexities and employing appropriate visualization techniques, the study contributes to refining skills in data visualization and fosters a deeper comprehension of complex datasets.

V. Conclusion

Data visualization offers multifaceted benefits across various domains. Firstly, it facilitates clearer communication of complex data, making it easier for everyone to grasp insights and trends. Secondly, visual representations enable quick identification of patterns, outliers, and correlations, aiding in data-driven decision-making processes. Also, data visualization enhances storytelling by providing compelling visuals that engage audiences and convey key messages effectively.

On the other hand, the study on data visualization has the potential to significantly aid various sectors by simplifying complex data and enabling informed decision-making. In healthcare, it can help healthcare professionals better understand trends in drug effectiveness and patient satisfaction, ultimately improving patient care. Similarly, in business, research can assist organizations in optimizing operations and identifying market trends through clear data visualization. Additionally, in academic and research contexts, the study promotes collaboration and deeper insights into data dynamics. Overall, the findings serve as a valuable resource for leveraging data visualization to drive innovation and achieve impactful outcomes across different fields.

VI. References

[1]. M. Ribeiro, L. Azevedo, and M. J. Pereira (2024). EpiGeostats: An R Package to Facilitate Visualization of Geostatistical Disease Risk Maps12. Mathematical Geosciences, 56, 103-119. DOI: 10.1007/s11004-023-10080-y3.

[2]. Y. Rimal, S. Gochhait, and A. Bisht (2021). Data Interpretation and Visualization of COVID-19 Cases Using R Programming. Informatics in Medicine Unlocked, 26, 100705. DOI: 10.1016/j.imu.2021.100705.

[3]. E. Nordmann, P. McAleer, W. Toivo, H. Paterson, and L.M. DeBruine (2022). Data Visualization Using R for Researchers Who Do Not Use R. Advances in Methods and Practices in Psychological Science, 5(2), 1–36. DOI: 10.1177/25152459221074654.

[4]. A. T. Siddiqui (2021). Data Visualization: A Study of Tools and Challenges. Asian Journal of Technology & Management Research, 11(01). Retrieved from https://www.researchgate.net/publication/352735133

[5]. W. Yang, T. Ye, L. Tie-shan, and P. Dongcheng (2018). Visualization Analysis of Shipping Recruitment Information Based on R. Proceedings of the 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI), Xiamen, China, March 29–31, 2018.

[6]. D. Srivastava (2023). An Introduction to Data Visualization Tools and Techniques in Various Domains. International Journal of Computer Trends and Technology, 71(4), 125-130. DOI: 10.14445/22312803/IJCTT-V71I4P116

[7]. S. K. Ahammad Fahad (2018). Big Data Visualization: Allotting by R and Python with GUI Tools1. International Conference on Smart Computing and Electronic Enterprise (ICSCEE2018).

[8]. E. Nordmann, P. McAleer, W. Toivo, H. Paterson, and L. M. DeBruine (2022). Data Visualization Using R for Researchers Who Do Not Use R. Advances in Methods and Practices in Psychological Science, 5(2), 1–36. DOI: 10.1177/25152459221074654