Journal of Electronic Research and Application, 2024, Volume 8, Issue 3
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Comparison of R and Excel in the Field of Data Analysis

Jue Wang*

Syracuse University, New York 13244, USA

Corresponding author: Jue Wang, jwang144@outlook.com

Abstract: This research paper compares Excel and R language for data analysis and concludes that R language is more suitable for complex data analysis tasks. R language's open-source nature makes it accessible to everyone, and its powerful data management and analysis tools make it suitable for handling complex data analysis tasks. It is also highly customizable, allowing users to create custom functions and packages to meet their specific needs. Additionally, R language provides high reproducibility, making it easy to replicate and verify research results, and it has excellent collaboration capabilities, enabling multiple users to work on the same project simultaneously. These advantages make R language a more suitable choice for complex data analysis tasks, particularly in scientific research and business applications. The findings of this study will help people understand that R is not just a language that can handle more data than Excel and demonstrate that r is essential to the field of data analysis. At the same time, it will also help users and organizations make informed decisions regarding their data analysis needs and software preferences.

Keywords: Excel; R language; Data analysis; Open source; Compare; Data management; Advantages; Disadvantages; Function

Online publication: June 13, 2024

1. Introduction

Data analysis is important in many fields, from business and finance to healthcare and social sciences. Two popular tools for data analysis are Microsoft Excel and R language. Excel is a widely used spreadsheet software that allows users to perform basic data analysis tasks such as sorting, filtering, and basic statistical analysis. Excel is commonly used in finance, accounting, and other business-related fields to manage data and create reports. For example, in sales Excel can be used to track sales data, calculate commissions, and analyze sales trends [1]. On the other hand, R is a programming language used for statistical computing and graphics. It is open-source software that provides a wide range of statistical and graphical techniques for data analysis [2]. R is widely used in academic research, healthcare, and social sciences, among other fields. For example, in healthcare, R is used to analyze patient data and develop predictive models for diseases [3].

Data analysis is the process of examining data to extract meaningful insights and conclusions. It involves

various techniques such as data cleaning, visualization, and statistical analysis. An example is the specific application of data analysis in predicting climate change. Experts analyze data related to temperature, rainfall, and atmospheric conditions. This allows them to identify patterns and trends that help them understand how the climate is changing over time. Then, this collected information can be used to develop policies and strategies to mitigate the effects of climate change. Such as reducing greenhouse gas emissions and increasing the use of renewable energy. So, data analysis aims to make informed decisions based on the insights extracted from data. Furthermore, data analysis has become increasingly important in today's world due to the exponential growth of data generated by various sources such as social media, online transactions, and scientific research. The ability to analyze and interpret data effectively provides a competitive advantage in many fields. It helps businesses to make informed decisions, governments to develop policies, and researchers to discover new insights. As a result, the demand for professionals with skills in data analysis has grown tremendously. Learning data analysis skills can open up numerous career opportunities in a wide range of industries. In conclusion, data analysis is a crucial tool for understanding complex data and making informed decisions based on the insights gained.

In the context of data analysis, Excel and R language have different strengths and limitations. Excel is a useful tool for simple calculations and data manipulation tasks. However, it has limitations when it comes to handling large datasets and performing advanced statistical analysis. In contrast, R language offers numerous advantages over Excel. R can handle large datasets and is highly customizable, allowing users to create and modify functions to meet their specific needs. Moreover, R offers a wide range of statistical tools and techniques that are not available in Excel, making it a preferred tool for complex data analysis tasks.

Therefore, this paper aims to compare the applications of Excel and R language in data analysis, highlighting the unique advantages of R language. We will explore the differences in their capabilities, ease of use, and the types of analyses they can perform. We will also provide specific examples of R's applications in various fields to illustrate its usefulness in data analysis. Overall, our analysis will show that R language is an indispensable tool for data analysts and researchers, offering unmatched capabilities that Excel cannot match.

2. Literature review

Excel and R are the two most commonly used tools for data analysis. While Excel is widely used in business and academia, R has gained popularity as a powerful open-source tool for statistical analysis and data visualization. In Google Scholar, JSTOR, and ScienceDirect, the keywords searched include: "R," "data analysis," "statistics," "modeling," "simulation," "visualization," and "practical applications" to find essays on topics related to the characteristics and problems of the R language. Through "excel," "data analysis," "operability," "experimentation," and "transparency," we searched for essays on topics related to the characteristics and problems of the R language. The keywords "excel," "data analysis," "operability," "experimentation," and "transparency" were used to search for essays on the characteristics and problems of Excel in practice in specific projects. The literature search focused on articles published in the last 10 years.

This topic focuses on the advantages and disadvantages of using R and Excel in data analysis. It is more important and indispensable than Excel in the field of data analysis. In the article, the topic is divided into four directions: the advantages of R, the disadvantages of R, the advantages of Excel, and the disadvantages of Excel.

2.1. Advantages of the R language

In the document "Creating and Deploying an Application with (R) Excel and R" it is shown and explained how R can be used to extend the statistical analysis and data visualization capabilities of Excel. For example,

R can be used to perform advanced statistical analyses not available in Excel, such as multivariate analysis and machine learning. This means that the R language has greater flexibility and scalability to provide more and more advanced analysis methods for complex data analysis tasks that Excel cannot perform [4]. As well, R can also be used to create custom data visualization tools not available in Excel, such as interactive visualizations and heat maps. This gives data analysts more freedom to visualize data to better demonstrate the characteristics and patterns of data [5].

"Applied Spatial Data Analysis with R" discusses the use of R for spatial data analysis, highlighting the fact that R analysis visualizes spatial data (including GIS and remote sensing data) with key advantages. First of all, R has powerful data visualization capabilities in spatial analysis. R language provides two-dimensional and three-dimensional mapping and interactive graphics, which help users create detailed maps and visualizations [6]. Secondly, R's extensive statistical techniques include spatial autocorrelation, spatial regression, and spatial interpolation, which analyze complex spatial datasets and extract meaningful insights. This literature discusses custom functions and algorithms for spatial data analysis in the R language, and "Creating and Deploying" again highlights the flexibility and extensibility of the r language. Also, since the R language itself is in the form of code, R provides a highly reproducible environment for spatial data analysis, which is essential for scientific research. R's ability to produce reproducible results ensures that findings can be validated and replicated by other researchers [7-9].

"The R Project in Statistical Computing" presents the idea that R has a wide range of advanced statistical methods. The ability of the R language to perform survival analysis is emphasized, and it is used to model the time before an event occurs. This type of analysis is not available in Excel and is commonly used in medical research and other fields. This literature also provides an example of custom data visualization in R, showing how heat maps can be used to visualize gene expression data [10,11].

2.2. Disadvantages of the R language

In "R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics," the authors mention that the learning curve for R is steeper than for Excel. This may be a disadvantage for some users. R requires users to write code, which can be intimidating to those unfamiliar with programming [12]. Despite these potential drawbacks, its ability to perform advanced statistical analysis and create custom data visualizations makes it a valuable tool for data analysts [13].

2.3. Advantages of Excel

In their paper "R Through Excel," Heiberger and Neuwirth point out that Excel is widely used and known, and many users are already familiar with its interface and functionality. In addition, Excel is relatively easy to use, especially for simple data analysis and visualization, and provides a range of built-in functions and tools that allow users to quickly manipulate and visualize data without requiring extensive knowledge of coding [14]. For example, Excel offers a range of chart types and customization options that allow users to easily create simple data visualizations [13].

In "Creating and Deploying an Application," the authors likewise cite Excel's broad usability and user-friendliness. In addition, Baier points out that Excel has built-in tools and features that allow users to quickly perform basic data analysis tasks, and while the performance is not as advanced as R, it still can handle basic data analysis. such as filtering, sorting, and charting. The authors place special emphasis on Excel's charting capabilities, noting that Excel's chart types are easy to use and can help convey insights about data clearly and concisely [5].

2.4. Disadvantages of Excel

"It's easy to Produce Chartjunk" highlights several limitations and challenges associated with using Excel for data visualization. Although Excel offers a range of chart types and customization options, using the tool to create effective, content-rich visualizations can be difficult. One of the main limitations is the tendency for users to create "chartjunk," or cluttered and confusing visualizations that do not effectively communicate the underlying data. This can be a particular challenge with Excel, as it offers users a large selection of chart types, formats, and labels. In addition, Excel's default settings and formatting options can lead to misleading or ineffective visualizations. For example, Excel's default 3D chart type can distort data, making it difficult to interpret accurately. Similarly, default color schemes and formatting options may not be appropriate for all data types or user needs, especially for users with color vision deficiencies [15].

Guerrero points out in "Excel Data Analysis" that although Excel can be a useful tool for data analysis, modeling, and simulation. However, it has some limitations that may make it less effective than more advanced tools such as R. One of the main limitations is Excel's limited ability to handle large data sets. While Excel can handle medium-sized data sets, it can become slow or unstable when working with data sets containing thousands or millions of rows. This can limit the types of analyses and models that can be performed using Excel, especially for complex or computationally intensive tasks. In addition, Excel's data cleaning and preparation capabilities may be limited, and users may need to supplement Excel with other tools or techniques to effectively clean and prepare their data for analysis. Finally, Excel's built-in statistical and modeling capabilities may be limited compared to more advanced tools such as R. While Excel offers a range of statistical functions and chart types, these may not be sufficient to perform more complex modeling tasks or perform advanced statistical analysis [16].

3. Discussion

When using Excel, most of the work can be done with a mouse click and you can access various tools in different places within the interface. So, Excel is very easy to use, but working with data in Excel is very time-consuming and the processes have to be repeated monotonously if you take on a new project. When using R, all operations are done in code. The data is loaded into memory and then a script is run to study and process the data. This tool may not be user-friendly, but it offers several benefits. Conceptually, R is easier to use. When working with multiple columns of data, you can see all the data even though you are only working on a single task. With R, on the other hand, the data is all in memory and can only be seen by pulling it up. If a data transformation or calculation is being performed, the user works with a subset of the relevant columns or rows, while all other data remains in the background. This makes it easier to focus on the task at hand. After completing the task, it can be saved in a data frame that contains only the required column or row data. Once the user has built the correct data set, the current problem can be solved. This may seem insignificant, but it is actually very beneficial.

With R, the same operations can be easily repeated for other data sets. Since all data is processed and studied in code, it is easy to perform the same operation on a new data set. With Excel, most operations are performed with a mouse click, which is a good user experience, but it is time-consuming and tedious to repeat operations on new data. R simply loads a new data set and runs the script again.

In fact, working with code also makes it easier to diagnose and share analysis results. When using Excel, most of the analysis results are memory-based (pivot table here, formula editor on another table). In R, on the other hand, all operations are performed by code, at a glance. If an error is being fixed, users know exactly where to do it, and if they need to share analysis results, they can simply copy and paste the code. When looking

for help online, users can specify exactly what data they are using and ask specific questions. In fact, most of the time, when asking questions online, people just post the exact code to solve the problem.

Project organization in R is much simpler. In Excel, a series of tables and possibly multiple workbooks need to be prepared and named appropriately, and the file names must not be duplicated. Project notes are kept in separate files. Conversely, R project organization uses a separate folder where everything that has been processed is placed, including cleaned data, exploratory charts, and models. This structure makes it easy to understand and locate, providing convenience for others collaborating on the project. Any data can be loaded into R. Regardless of where or how the data is stored, R can load CSV files, read JSON, perform SQL queries, or extract data from websites. Even big data can be processed in R via Hadoop.

R is a complete toolset that uses data packages. R is more useful than Excel when it comes to analyzing data. R can be used to perform data management, classification, and regression, as well as to process images and perform all other operations. With over 5,000 packages currently available, R can handle any type of data [17]. One of the most useful features of R is its excellent data visualization. With ggplot2, you can quickly create as many charts as you need and adjust them to the shape of the chart itself. And ggplot2 can also create many more types of charts [18]. For example, it is not possible to create a scatter plot matrix with Excel, but R can create such a matrix very easily.

4. Conclusion

According to the above discussion in the literature, the advantages of the r language in the field of data analysis far outweigh the advantages of Excel. The only disadvantage of the human language is that it is difficult to learn, while Excel has a lot of hard conditions that are not enough (**Figure 1**). Therefore, the R language has absolute advantages over Excel in data analysis. In conclusion, this research paper compares data analysis in Excel and R language and proves that R language is more suitable for complex data analysis tasks. the open-source nature of R language makes it accessible to all and its powerful data management and analysis tools make it suitable for handling complex data analysis tasks. In addition, the R language offers high reproducibility capabilities, making it easy to replicate and validate findings, and it has excellent collaboration capabilities, enabling multiple users to work on the same project simultaneously (**Figure 1**). The results of this study underscore the importance of considering an organization's specific data analysis needs when choosing between Excel and the R language [19]. While Excel is a useful tool for simple data analysis tasks, it has limitations when working with large data sets and performing advanced statistical analysis. In contrast, R offers unparalleled capabilities for complex data analysis tasks and can be customized to meet specific needs. Overall, the comparison of Excel and R language in this research paper demonstrates the importance of using the right tools to get the job done. For simple data analysis tasks, Excel may be sufficient, but for more complex tasks, the R language is essential. This research paper is intended to help data analysts and researchers make informed decisions about their software preferences, and we hope it will contribute to advancing the field of data analysis.

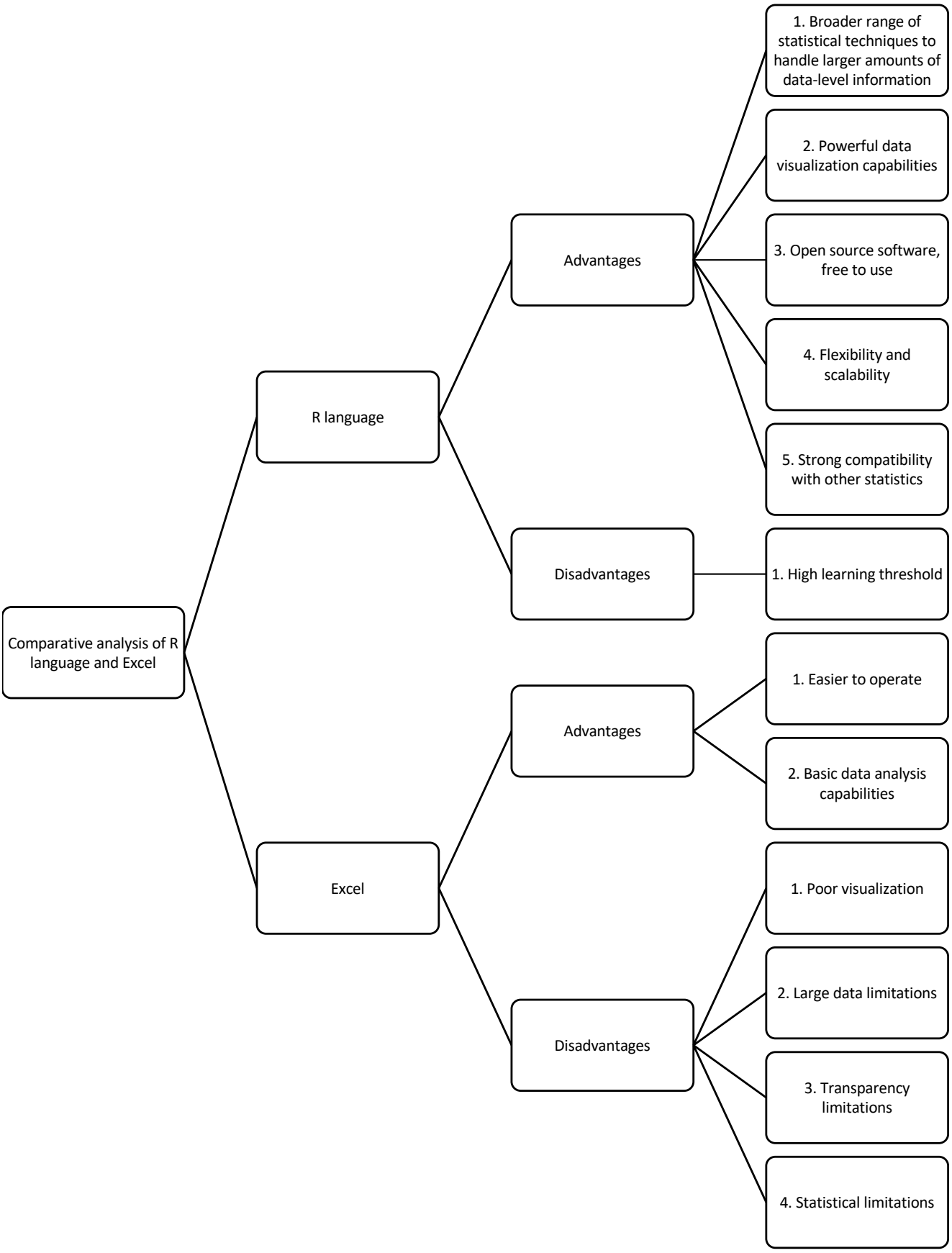

Figure 1. Comparative analysis of R language and Excel

Disclosure statement

The author declares no conflict of interest.

References

[1] Meyer DZ, Avery LM, 2008, Excel as a Qualitative Data Analysis Tool. Field Methods, 21(1): 91–112. https://doi.org/10.1177/1525822x08323985

[2] R Core Team, 2018, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, viewed April 26, 2023, https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing

[3] Grolemund G, Wickham H, 2016, R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, O'Reilly Media, Newton.

[4] Chambers JM, 2008, Programming with Data: A Guide to the S Language, Springer Science + Business Media, Berlin.

[5] Baier T, Neuwirth E, Meo MD, 2011, Creating and Deploying an Application with (R)Excel and R. The R Journal, 3(2): 5–11. https://doi.org/10.32614/rj-2011-011

[6] Murrell P, 2011, R Graphics, CRC Press, Boca Raton.

[7] Su YS, 2008, It's Easy to Produce Chartjunk Using Microsoft® Excel 2007 but Hard to Make Good Graphs. Computational Statistics & Data Analysis, 52(10): 4594–4601. https://doi.org/10.1016/j.csda.2008.03.007

[8] Wild CJ, Pfannkuch M, 2012, The Importance of Statistical Thinking in Research. American Psychologist, 67(4): 280.

[9] Altman N, Krzywinski M, 2015, Points of Significance: Reproducibility. Nature Methods, 12(9): 831.

[10] Ripley BD, 2001, The R project in Statistical Computing. MSOR Connections, 1(1): 23–25. https://doi.org/10.11120/msor.2001.01010023

[11] Gentleman R, Carey VJ, Bates DM, et al., 2004, Bioconductor: Open Software Development for Computational Biology and Bioinformatics. Genome Biology, 5(10): R80.

[12] Burns P, 2011, The R inferno, Lulu.com.

[13] Heiberger RM, Neuwirth E, 2009, R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics (1st. Ed), Springer, Berlin.

[14] Robbins D, 2011, Excel 2010 Power Programming with VBA, John Wiley & Sons, Hoboken.

[15] Bivand RS, Pebesma E, Gómez-Rubio V, 2013, Applied Spatial Data Analysis with R, Springer, Berlin.

[16] Guerrero H, 2019, Excel Data Analysis: Modeling and Simulation, Springer, Berlin.

[17] Kuhn M, 2008, Building Predictive Models in R Using the Caret Package. Journal of Statistical Software, 28(5): 1–26.

[18] Chang W, 2019, R Language for Data Science, CRC Press, Boca Raton.

[19] Walther A, 2018, Excel vs. R: The Battle for Data Analysis Supremacy, Medium.