

## OPEN-SOURCE SOFTWARE FOR ADVANCED STATISTICAL DATA ANALYSIS

Yogesh Garde, Nitin Varshney\*, and Alok Shrivastava

Department of Agricultural Statistics, Navsari Agricultural University, Navsari, Gujarat– 396450

\*Corresponding author email: [nitin.caw@nau.in](mailto:nitin.caw@nau.in)

### ABSTRACT

The utilization of open-source software has significantly revolutionized the landscape of advanced statistical data analysis. This paper explores the dynamic role that open-source software plays in empowering researchers, analysts, and organizations to conduct sophisticated statistical analyses. By offering accessible tools and resources, open-source platforms have democratized data analysis, fostering collaboration, innovation and cost-effectiveness. Through a comprehensive investigation of popular open-source software alternatives, this paper highlights concise applications in various fields. From enhancing research methodologies to enabling evidence-based decision making, the paradigm shift toward open-source software is reshaping the way we approach and use statistical insights for multiple purposes.



### INTRODUCTION

Open-source software for statistical data analysis has become increasingly important and prevalent in recent years. This article highlights the significance and the recent era of usage of open-source software in statistical data analysis. Open-source software provides numerous advantages, including accessibility, flexibility, transparency, and cost-effectiveness. These features have made it a popular choice among researchers, analysts, and organizations in various fields. It breaks down barriers to access and allows users to freely obtain, use, modify, and distribute the software. Open-source software also provides a wide range of statistical techniques, algorithms, and visualization tools that cater to diverse analytical needs.

In the recent era, open-source software has gained significant traction due to several factors:

- The exponential growth of data requires scalable and efficient tools for analysis. Open-source software enables the processing and analysis of large and complex datasets. This is crucial in fields such as big data analytics, bioinformatics, social sciences, finance, etc.
- The advancements in machine learning and artificial intelligence have increased the demand for open-source software that supports these techniques.
- The growing emphasis on reproducibility and transparency in research has fuelled the adoption of open-source software. Researchers are increasingly turning to open-source tools for their data

analysis needs, as it allows them to share their code, methods, and results openly, facilitating the replication and validation of their findings.

- d. The cost-effectiveness of open-source software makes it an attractive option for organizations and individuals.

There are several open-source software options available for statistical data analysis. A few of them are listed below:

**1. R & R STUDIO:**

R is a widely used programming language and software environment for statistical computing and graphics. It provides a vast collection of packages for various statistical analysis tasks and data visualization. RStudio is a popular integrated development environment (IDE) for working with R.



**Applications of R:**

| Sl No. | Application                         | Description  | Packages  |
|--------|-------------------------------------|--|---|
| 1      | Data Analysis and Exploration       | R is extensively used for exploratory data analysis  | R's data manipulation packages like <i>dplyr</i> and <i>tidyr</i> make it easy to manipulate and reshape data for analysis.   |
| 2      | Statistical Modelling               | R is a powerful tool for statistical modelling and inference   | <i>stats</i> , <i>lme4</i> , <i>survival</i> , and <i>brms</i> .  |
| 3      | Data Visualization                  | R offers numerous options for creating high-quality visualizations   | <i>ggplot2</i>  |
| 4      | Machine Learning                    | R provides a rich set of packages for machine learning tasks   | <i>Caret</i> , <i>random Forest</i> , <i>xgboost</i> , and <i>keras</i>   |
| 5      | Bioinformatics and Genomics         | R is widely used in the field of bioinformatics for analysing and interpreting genomic data                              | <i>Bioconductor</i>   |
| 6      | Econometrics and Finance            | R is commonly used in econometrics and finance for analysing economic data and financial markets                         | <i>plm</i> for panel data analysis<br><i>forecast</i> for time series forecasting<br><i>quantmod</i> for financial modelling<br><i>Portfolio Analytics</i> for portfolio optimization |
| 7      | Social Sciences                     | R is popular among researchers in social sciences, including psychology, sociology, and political science                | <i>plyr</i>   |
| 8      | Data Science and Big Data Analytics | R is frequently used in data science workflows, combining data manipulation, statistical modelling, and machine learning | <i>sparklyr</i> and <i>dplyr</i>  |

**Some popular sources to help you get started with learning R:**

- i. **R Documentation:** The official R documentation (<https://www.r-project.org/>) is a comprehensive resource that provides detailed information about R's functions, packages, and syntax.

- ii. **RStudio Education:** RStudio, the popular IDE for R, offers a dedicated website for learning R called RStudio Education <https://education.rstudio.com/>
- iii. **RStudio Online Learning:** <https://rstudio.com/online-learning/>
- iv. **R for Data Science:** "R for Data Science" by Hadley Wickham and Garrett Grolemund is a widely recommended book for learning R.
- v. **DataCamp:** It is an online learning platform that offers interactive courses on R and data science <https://www.datacamp.com/>
- vi. **Coursera:** Coursera (<https://www.coursera.org/>) offers several R-related courses, including "R Programming" by Johns Hopkins University and "Data Science and Machine Learning Bootcamp with R" by the University of Washington.
- vii. **YouTube Tutorials:** YouTube has a wealth of R tutorials and channels dedicated to R programming. Channels like "R Programming A-Z" and "R Tutorial Series" provide comprehensive video tutorials on various aspects of R, from the basics to advanced topics.
- viii. **R-Bloggers:** R-Bloggers (<https://www.r-bloggers.com/>) is a popular community-driven blog aggregator that compiles R-related articles, tutorials, and resources from various sources.
- ix. **Stack Overflow:** Stack Overflow (<https://stackoverflow.com/>) is a question-and-answer website where you can find answers to specific R programming questions. There are numerous books available that provide in-depth coverage of R programming and statistical analysis. Some popular titles include "*R for Data Science*"  
by Hadley Wickham and Garrett Grolemund, "*The Art of R Programming*" by Norman Matloff, and "*R Cookbook*" by Paul Teetor.

## 2. PYTHON:

Python is a versatile programming language that offers various libraries for statistical analysis. *NumPy*, *SciPy*, and *pandas* are commonly used libraries for numerical computing and data manipulation. Additionally, libraries like *StatsModels* and *scikit-learn* provide statistical modelling and machine learning capabilities.



**Applications of Python:**

| SI No. | Application                                  | Description  | Libraries   |
|--------|--|--|---|
| 1      | Data Analysis and Exploration                | Python is widely used for data analysis and visualization tasks  | <i>NumPy, pandas, matplotlib, Seaborn and Plotly</i>                                    |
| 2      | Scripting and Automation                     | for writing scripts and automating tasks   | <i>stats, lme4, survival, and brms.</i>   |
| 3      | Web Development                              | Python has several frameworks, which simplify web development  | <i>Django and Flask</i>   |
| 4      | Machine Learning and Artificial Intelligence | It is a new choice for many machine learning and AI applications   | <i>scikit-learn, TensorFlow, PyTorch, and Keras</i>                                     |
| 6      | Scientific Computing                         | Python is widely used in scientific computing and simulation   | <i>SciPy</i>  |
| 7      | Internet of Things (IoT)                     | Python is used in IoT projects for data collection, device communication, and analytics                                  | <i>PySerial</i> , frameworks like Raspberry Pi and MicroPython support IoT development. |
| 8      | Natural Language Processing (NLP)            | R is frequently used in data science workflows, combining data manipulation, statistical modelling, and machine learning | NLTK (Natural Language Toolkit), <i>spaCy</i>   |

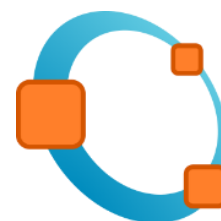
**Some popular sources to help you get started with learning Python:**

- i. **Python's Official Documentation:** Python's official website (<https://www.python.org/>) provides comprehensive documentation that covers all aspects of the language. It includes tutorials, guides, and references for both beginners and advanced users. The "Python Tutorial" section is an excellent starting point.
- ii. **Codecademy:** Codecademy offers an interactive Python course for beginners that allows you to learn Python syntax and concepts through hands-on coding exercises. You can access it at <https://www.codecademy.com/learn/learn-python>.
- iii. **Automate the Boring Stuff with Python:** This popular online book by Al Sweigart is aimed at beginners who want to learn Python by automating everyday tasks. The book is available for free at <https://automatetheboringstuff.com/>.
- iv. **Python Crash Course:** Written by Eric Matthes, Python Crash Course is a beginner friendly book. The book is available at <https://nostarch.com/pythoncrashcourse>.
- v. **Real Python:** Real Python is an online platform that offers tutorials, articles, and video courses on various Python topics. You can find their resources at <https://realpython.com/>.
- vi. **"Learn Python the Hard Way" (LPTHW):** This book by Zed Shaw provides a hands-on approach to learning Python. You can access the Python 3 version at <https://learnpythonthehardway.org/python3/>.

- vii. **Python.org Beginner's Guide:** Python.org provides a Beginner's Guide that offers step-by-step instructions to learn Python. You can find it at <https://docs.python.org/3/tutorial/index.html>.
- viii. **YouTube Tutorials:** YouTube hosts numerous Python tutorial channels that cater to different learning styles. Channels like Corey Schafer, Sentdex, and Tech with Tim offer beginner-friendly Python tutorials and cover various Python topics with practical examples.

### 3. GNU OCTAVE:

GNU Octave is a high-level programming language primarily intended for numerical computations and scientific simulations. It is compatible with MATLAB syntax and provides a wide range of built-in functions for statistical analysis and data visualization. <https://www.gnu.org/software/octave/>



### 4. JUPYTER NOTEBOOK:

Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text. It supports multiple programming languages, including R and Python, making it an excellent choice for interactive data analysis.



#### Some popular sources to help you get started with learning Jupyter Notebook:

- i. **Jupyter Notebook Documentation:** The official Jupyter Notebook documentation <https://jupyter-notebook.readthedocs.io/>.
- ii. **Jupyter Notebook Tutorial on Real Python:** Real Python offers a comprehensive tutorial on Jupyter Notebook (<https://realpython.com/jupyter-notebook-introduction/>)
- iii. **YouTube Tutorials:** Many YouTube channels offer Jupyter Notebook tutorials for different levels of expertise. Channels like Corey Schafer, DataCamp, and Jupyter offer video tutorials.
- iv. **Jupyter Notebook Examples on GitHub:** The Jupyter organization on GitHub (<https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks>).

### 5. JULIA:

Julia is a high-level programming language specifically designed for numerical and scientific computing. It offers a comprehensive ecosystem of packages for statistical analysis, data visualization, and machine learning. It is available on <https://juliaacademy.com/>.



### 6. KNIME:

KNIME (Konstanz Information Miner) is an open-source data analytics platform that allows you to visually create data workflows, combining different data processing and analysis steps.



It supports a wide range of statistical analysis techniques and integrates with various programming languages and tools. It is available on <https://hub.knime.com/>, <https://www.knime.com/learning-hub>.

## 7. ORANGE:

Orange is an open-source data visualization and analysis tool that provides a visual programming interface for building data analysis workflows. It offers a wide range of statistical and machine learning methods and supports interactive data visualization. It is available on <https://docs.orange.biolab.si/>.



## 8. PSPP:

PSPP is a free and open-source alternative to IBM SPSS, a popular commercial software for statistical analysis. PSPP provides a user-friendly interface and supports basic statistical procedures, including descriptive statistics, hypothesis testing, and regression analysis. <https://www.gnu.org/software/pspp/manual/>



## 9. APACHE SPARK:

Spark is an open-source big data processing framework that includes a module called MLlib for scalable machine learning and statistical analysis. It provides a distributed computing environment and supports various statistical algorithms. It is freely available on <https://spark.apache.org/documentation.html>.



## 10. OPSTAT

Prof. O.P. Sheoron, Statistics programmer, Chaudhary Charan Singh Haryana Agricultural University developed web based statistical software for data analysis. (<http://14.139.232.166/opstat/>)

## 11. PBTOOLS:

Plant Breeding Tools (PBTools) is a software that has been developed to assist plant breeders in the design of experiments and analysis of data. It has an easy-to-navigate graphical user interface that does not require users to have programming skills to perform data manipulation and analysis. It uses R functions that were specifically written for the development of this software.



## 12. STATISTICAL TOOL FOR AGRICULTURAL RESEARCH







### (STAR):

STAR is a **computer** program for data management and basic statistical analysis of experimental data. It has a user-friendly graphical interface where items are accessible via drop-down menus. Its graphical interface was created using the Eclipse Rich Client Platform (RCP) and uses the R language and environment for statistical computing and graphics. The program uses functions in R that are specifically written for the development of this computer program. STAR has been developed primarily for the analysis of data from agricultural field trials, but many of the features can be used for analysis of data from other sources.



### Some other open-source software which can be used in data analysis are listed below:

| Sr. no | Name of the Software | Whether Open-source | Free version available | Best Feature                                      |
|--------|----------------------|---------------------|------------------------|---|
| 1.     | Atom                 | Yes                 | ✓                      | Key binding customization                         |
| 2.     | Brackets             | Yes                 | ✓                      | Pre-processor support                             |
| 3.     | Bluefish             | Yes                 | ✓                      | Can open 500+ files at a time                     |
| 4.     | Visual Studio code   | Yes                 | ✓                      | Intellisense function                             |
| 5.     | Notepad++            | Yes                 | ✓                      | Plug-in for MIME tools, NPP export, and converter |
| 6.     | Cuda text            | Yes                 | ✓                      | Smart auto-completion for HTML, CSS               |
| 7.     | Emacs                | Yes                 | ✓                      | Debugger interface                                |
| 8.     | ConText              | No                  | ✓                      | Powerful command-line handler                     |
| 9.     | Editpadlite          | No                  | ✓                      | Opens large text files with ease                  |
| 10.    | Komodo IDE           | Yes                 | ✓                      | Code folding and cold blocks                      |

|   |   |
|---|---|
| AGD-R (Analysis of Genetic Designs with R)                    |    |
| META-R: Multi Environment Trial Analysis with R               |   |
| Molecular Evolutionary Genetics Analysis (MEGA)               |  |
| Trait Analysis by association, Evolution and Linkage (Tassel) |  |
| DARwin: Dissimilarity analysis and Representation for Windows |  |
| PAST  |  |

## SUMMARY

In conclusion, the integration of open-source software has brought about a substantial transformation in the realm of advanced statistical data analysis. This study delves into the dynamic role of open-source software in empowering researchers, analysts, and organizations to conduct intricate statistical analyses. These accessible platforms have democratized data analysis, promoting collaboration, innovation, and efficiency. Through a thorough exploration of prevalent open-source software options, this paper succinctly illustrates their diverse applications across various domains. From refining research methodologies to facilitating evidence-driven decision-making, the shift towards open-source software is reshaping the utilization of statistical insights across multiple objectives.

## REFERENCES

Open Source (Free) Statistical Software. Retrieved from: <https://u.osu.edu/qmc/open-source/>



Quantitative Analysis Software- Open Source. Retrieved from: <https://researchguides.uoregon.edu/How-to-Choose-Data-Analysis-Software/open>

The 7 Data Analysis Software Applications You Need to Know. Retrieved from: <https://www.coursera.org/articles/data-analysis-software>

\*\*\*\*\*