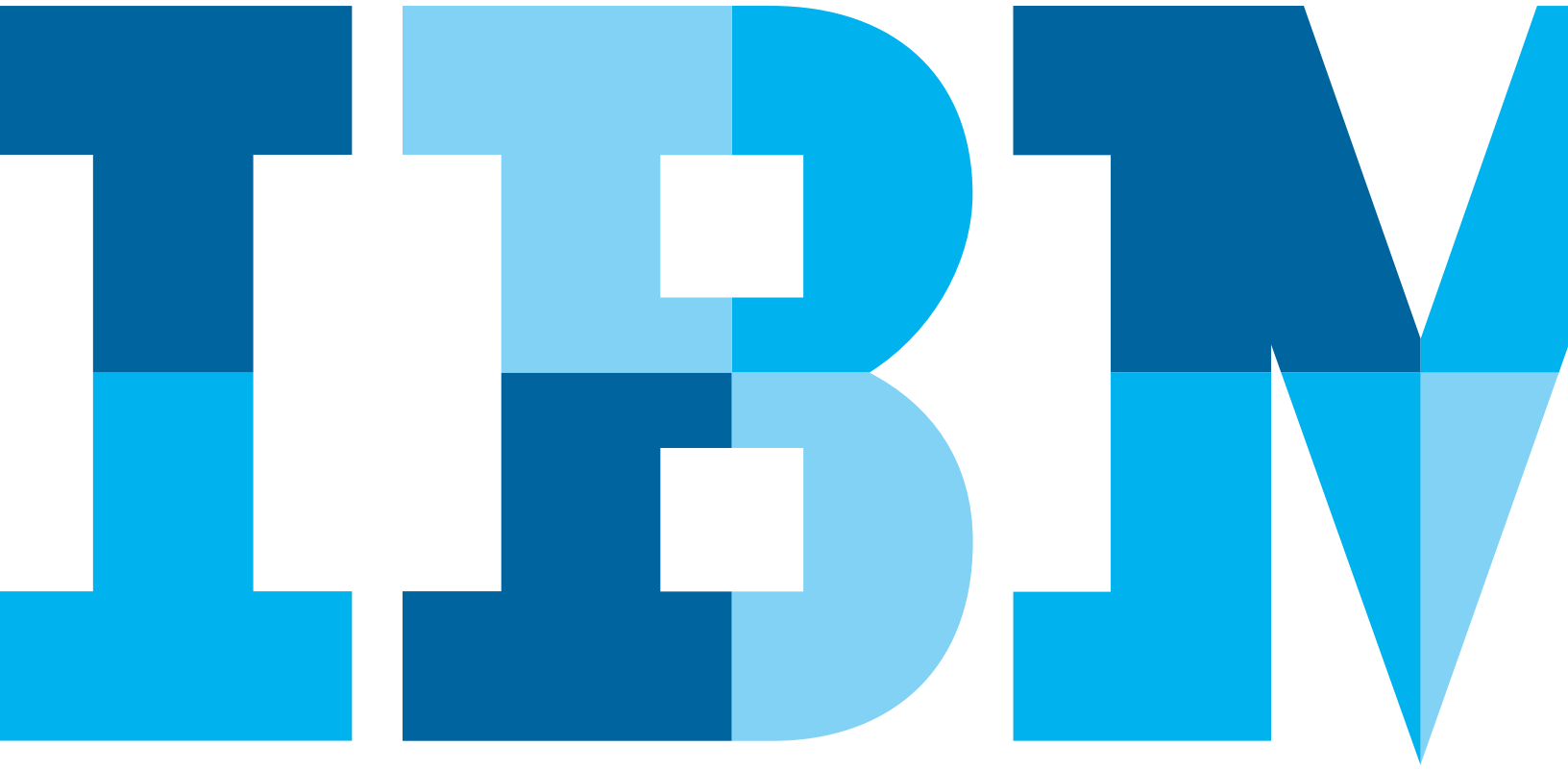
# R and SPSS software: Everyone wins



IBM

## Contents



## Overview

The purpose of this paper is to demonstrate the benefits of using the R programming language with IBM® SPSS® Statistics and Modeler software rather than simply trying to go it alone with R. With SPSS software, R users get access to superior data management, a point-and-click interface, presentation-quality output and improved scalability. SPSS software users get access to a rich, ever-expanding collection of statistical analysis and graphing libraries to help them gain deeper insights from their data.

Using SPSS software and R together makes the most of both worlds. SPSS software can run R syntax from your SPSS interface. You can augment the powerful data manipulation, statistical analysis and predictive algorithms already in SPSS software with custom R code for additional power and flexibility. You can conduct custom analysis, create and work with output and integrate new insights into your analysis plans. In addition, with the SPSS Custom Dialog Builder, you can share and reuse R code with those who would benefit from the new analysis options but choose not to use programming for analysis. Developers can focus on writing code while users can focus on analysis and use R-based functionality without learning R.

## What is R?

R is an open source programming language and software environment for statistical computing and graphics (www.r-project.org). The R language has become very popular with statisticians and data miners who use it to develop statistical software. In addition, R is widely used for advanced data analysis. R provides a wide variety of statistical and graphical techniques such as linear and nonlinear modeling, classical statistical tests, time-series analysis, classification and clustering. R is available under the terms of the Free Software Foundation and GNU General Public License.

R has more than 4800 packages available from multiple repositories that specialize in econometrics, data mining, spatial analysis, bioinformatics and many more. R has been reviewed by internally renowned statisticians and computational scientists. However, because it is open source, releases and packages have no codified processes.

## The limitations of R

Because R is available at no charge, a common perception is that it can be used in place of commercial statistics and modeling software at great savings to an organization. This perception is somewhat misguided. Despite its considerable merits, R cannot offer everything you need to derive the most business value from your analytics. The following limitations are topics of concern for R users:

- *Deployment.* Using R to integrate predictive outputs into an operational environment can be difficult.
- *Interface.* R does not have a modern graphical user interface, which makes it difficult for those who are not R programmers to use it.
- *Learning curve.* R is not easy to learn for everyone. Not everyone is a programmer.
- *Data.* R does not easily connect to databases natively.
- *Output.* Production of publish-ready output is difficult.
- *Performance.* R can very quickly consume all available memory.
- *Collaboration.* R makes sharing work among an analyst team difficult, especially when team members do not have the same level of R knowledge.
- *Enterprise security.* The security of the packages that you download is not assured.

In addition, users must ensure that they have the right R packages and the code to string the packages and output together. These limitations often add costs to using R.

## Benefits of integrating R with SPSS software

Addressing the limitations of R by using it from within SPSS software makes sense. The combined strength of both helps address the needs of an organization that has few experts and that wants to benefit from R without a steep learning curve. SPSS software is a convenient platform from which R users can handle large data sets and get high quality graphs and other forms of output. Some of the other benefits are the ease of use of SPSS software and the ability to distribute integrated R packages to a wide range of users who are not familiar with R. Most importantly, when you use R from SPSS software, the limitations of R can be addressed.

### Deployment

SPSS software can cost-effectively handle the deployment limitations of R. For example, IBM SPSS Modeler Gold enables users to optimize how business rules and predictive models affect sales, customer service, maintenance and more. SPSS Modeler Gold supports the use of models that incorporate R. SPSS Modeler also streamlines access to business intelligence through integration with IBM Cognos® TM1®. IBM is able to provide users with multiple ways of using predictive intelligence. Users can integrate predictive insight into their interactive, mobile dashboards alongside historical and real-time data views without the need to create or purchase additional software.

### Interface

SPSS Modeler and SPSS Statistics provide a simple graphical user interface that supports a variety of data preparation, statistical analysis and predictive modeling algorithms. R code runs in the same interface, alongside all of the other functions and features that are already provided. Users can also add a GUI front-end to R. With this front-end, users who are not programmers can pass in customer values (Figure 1), such as variable names and more. They are able to take advantage of advanced analysis without having to understand the code underneath.

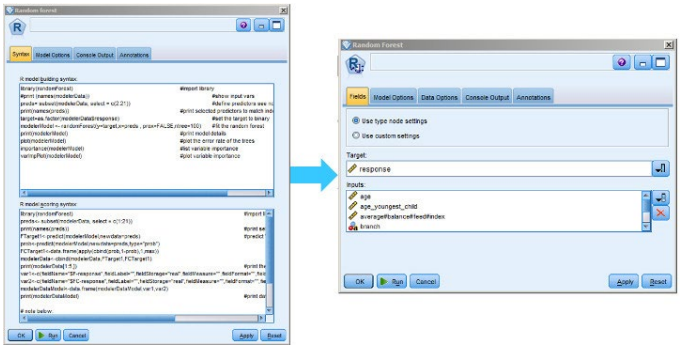


*Figure 1*: R code can be used to create customer dialogue boxes, making it easier for non-programmers to take advantage of the code during analysis.

## Learning curve

R is a programming language and therefore, it is something that must be learned. Yet, learning R is no easy task. And, while you are taking the time to learn R, you are not able to take advantage of its full capabilities. However, when R is integrated with SPSS software, you can exploit its benefits almost right away. You can focus your learning on those R routines that are truly unique and use the data, statistical analysis and modeling that is already included with SPSS software.

## Data

With R, accessing the data needed for analysis requires a great deal of time and effort. You must write volumes of code, implement packages and even employ Java. SPSS Statistics and SPSS Modeler remove the problem of data access. Both can be used with SQL, Oracle, SAP, IBM Netezza®, DB2® commercial databases and more. SPSS Statistics and SPSS Modeler can read text input, spreadsheets, SAS files and other formats. SPSS Modeler can read directly from IBM Cognos Business Intelligence and Cognos TM11. Wizards with prebuilt connectors access the data, which removes the excessive time-consuming burden of extracting, transforming and manipulating data before analysis. SPSS Statistics and SPSS Modeler provide powerful data manipulation techniques, encapsulated in point and click interfaces. Users can transpose, check and reformat data. SPSS Modeler also provides automatic data preparation, which optimizes data for predictive modeling with a single click.

## Output

SPSS Statistics and Modeler include multiple means of producing presentation-ready charts and graphs. SPSS Statistics software enables R programmers to wrap R functions in SPSS software syntax. As a result, you can produce presentation-quality graphs, pivot tables and other forms of output (Figure 2). Users can also easily publish their results to popular formats such as PDF, Word, PowerPoint, Excel and more. In addition, you can integrate R with SPSS Modeler advanced capabilities such as entity analytics, social media analytics, text analytics and more.

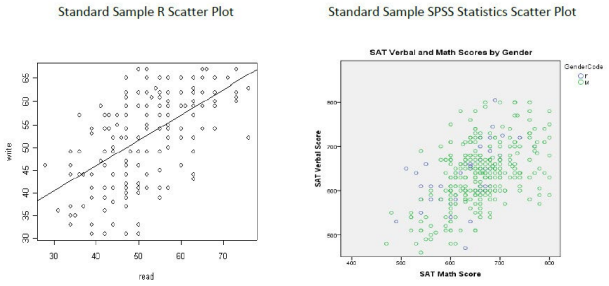



*Figure 2*: SPSS produces presentation-ready charts and graphs, which saves the time that business stakeholders often spend preparing and formatting analytical output.

## Performance

R commands operate in an in-memory workspace. R commands are not designed to take memory into account, and during an R session, all objects are stored in a temporary, working memory. As a result, available memory can be used up fairly quickly. The combination of R with SPSS Modeler enables you to partition or sample the data passed to R (Figure 3). In addition, SPSS Modeler Server is a memory-exploiting technology that can spill analysis over to disk so memory remains available. You can use R commands and create R objects without a major impact on the overall performance of your data mining and modeling. Also, SPSS Modeler can scale your R in-database for IBM Netezza and SAP Hana environments, among others. And, with SPSS Analytic Server, it can scale R in Hadoop as well.

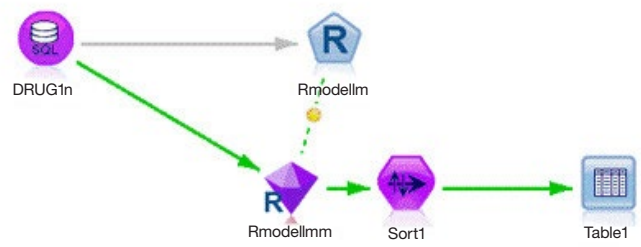



*Figure 3*: IBM SPSS Modeler is used to natively access data in an SQL database with an R model used to score records. Purple indicates that the analytical step is using SQL pushback, which improves performance by executing memory-intensive tasks in the database itself.

## Collaboration

R is a language that is designed for the individual. However, most analytics work is a collaborative effort with a number of people contributing to models and statistical analysis. When you use SPSS software and R together, you lose the lone wolf aspect of the language and gain IBM's world-class collaboration capabilities. SPSS software provides a complete framework for centralizing, securing and automating analytical assets developed with SPSS Statistics and SPSS Modeler. This ensures that models and statistical analysis developed can be shared securely and corporate governance can be applied.

## Enterprise security

R lacks a formal release process, which means formal quality assurance is not a part of it. Without QA, you are using R at your own risk. You cannot be sure that the release or package will provide the functions it claims to deliver. In addition, R packages are downloaded from the web and user communities so the security of the packages can be questionable. In fact, without your knowledge, a download could introduce malware, a Trojan horse virus, data taps and more. SPSS Modeler and Statistics, on the other hand, are tested rigorously as part of IBM's software QA process. Because the software is from IBM, you do not have to use risky practices that can threaten the security of your environment.

## Conclusion

Both SPSS and R independently boast strengths that have been tested over time and are strongly accepted in the analytical community. Moreover, these strengths complement each other to create an even more powerful set of functions and features that benefit the analytical community as a whole. R users can access superior data management capabilities, which enable them to handle much larger data sets. Also, SPSS software provides R users with a richer set of graphical and pivot table output options, which can lead to a better user experience. Finally, SPSS software acts as an ideal deployment vehicle for distributing R packages to a wide range of users.

SPSS software users gain access to many more statistical functions. They can then carry out complicated analysis without the hassles of learning a complex programming language such as R. The advantages of using R and SPSS software together are very much worth considering.

## About Business Analytics

IBM Business Analytics software delivers data-driven insights that help organizations work smarter and outperform their peers. This comprehensive portfolio includes solutions for business intelligence, predictive analytics and decision management, performance management and risk management. Business Analytics solutions enable companies to identify and visualize trends and patterns in such areas as customer analytics that can have a profound effect on business performance. They can compare scenarios; anticipate potential threats and opportunities; better plan, budget and forecast resources; balance risks against expected returns and work to meet regulatory requirements. By making analytics widely available, organizations can align tactical and strategic decision making to achieve business goals. For more information, see **ibm.com**/business-analytics.

## Request a call

To request a call or to ask a question, go to **ibm.com**/business-analytics/contactus. An IBM representative will respond to your inquiry within two business days.



IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
March 2014



YTW03375-USEN-00