**r4stats.com**
*Analyzing the World of Analytics*

# The Popularity of Data Analysis Software

*by Robert A. Muenchen*

**Abstract:** This article presents various ways of measuring the popularity or market share of software for advanced analytics software including: Alpine, Alteryx, Angoss, C / C++ / C#, BMDP, FICO, IBM SPSS Statistics, IBM SPSS Modeler, InfoCentricity Xeno, Java, JMP, KNIME, Lavastorm, Mathworks' MATLAB, Megaputer's PolyAnalyst, Minitab, NCSS, Python, R, RapidMiner, SAS, SAS Enterprise Miner, Salford Predictive Modeler (SPM) etc., SAP KXEN, TIBCO Spotfire, Stata, Statistica, Systat, WEKA / Pentaho.

I don't attempt to differentiate among variants of languages such as R vs. Revolution R Enterprise or TIBCO Enterprise Runtime for R; or SAS vs. the World Programming System (WPS) or Carolina, except when it is particularly easy such as comparing the company Pagerank figures.

Excluded from the list are products that focus on report writing (e.g. Cognos), or are tied to a specific database (e.g. Microsoft, Oracle, SAP's HANA), specific hardware (e.g. Teradata, IBM PureData) or a specific application field. I also exclude packages devoted to visualization, such as Tableau, Origin, or SigmaPlot.

The first section on jobs is currently the newest and covers the most extensive set of software. I'll add more software to the later sections as soon I can and announce the changes on Twitter where you can follow me as @BobMuenchen.

### Introduction

When choosing a tool for data analysis, now more commonly referred to as analytics, there are many factors to consider. Does it run natively on your computer? Does the software provide all the methods you use? If not, how extensible is it? Does that extensibility use its own language, or an external one (e.g. Python, R) that is commonly accessible from many packages? Does it fully support the style (programming vs. point-and-click) that you like? Are its visualization options (e.g. static vs. interactive) adequate for your problems? Does it provide output in the form you prefer (e.g. cut & paste into a word processor vs. LaTeX integration)? Does it handle large enough data sets? Do your colleagues use it so you can easily share data and programs? Can you afford it?

There are many ways to measure popularity or market share and each has its advantages and disadvantages. Here they are, in approximate order of usefulness:

- **Job Advertisements** – these are rich in information and are backed by money so they are perhaps the best measure of how popular each software is now, and what the trends are up to this point.
- **Scholarly Articles** – these are also rich in information and backed by significant amounts of effort. Since a large proportion come out of academia, the source of new college graduates, they are perhaps the best measurement of new trends in analytics.
- **Books** – the number of books that include a software's name in its title is a particularly useful information

since it requires a significant effort to write one and publishers do their own study of market share before taking the risk of publishing. However, it can be difficult to do searches to find books that use general-purpose languages which also focus only on analytics.

- **Website Popularity** – the PageRank measure is objective data, and for sites that clearly focus on analytics, it's unbiased and especially useful for weeding out the weaker software. However, so much market consolidation has occurred that now focused analytic tools like SPSS are listed under corporations with much broader interests (IBM in that case). In addition, for general-purpose software like Java, many sites that discuss programming point to http://www.java.com, that have nothing to do with its use for analytics.
- **Blogs** – the number of bloggers writing about analytics software is an interesting measure. Blog posts contain a great deal of information about their topic, and although it's not as time consuming as a book to write, maintaining a blog certainly requires effort. Unfortunately, this measure is very hard to collect except where sites exist to maintain such lists.
- **Surveys of Use** – these add additional perspective, but they are commonly done using "snowball sampling" in which the survey taker tries to widely distribute the link and then vendors vie to see who can get the most of their users to participate. So long as they all do so with equal effect, the results can be useful. However, the information is often limited, because the questions are short and precise (e.g. "tools data mining" or "program languages for data mining") and responding requires just a few mouse clicks, rather than the commitment required to place a job advertisement or publish a scholarly article, book or blog post. As a result, it's not unusual to see market share jump 100% or drop 50% in a single year, which is *very* unlikely to reflect changes in actual use.
- **Discussion Forum Activity** – these web sites or email-based discussion lists can be a very useful source of information because so many people participate, generating many tens of thousands of questions, answers and other commentary for popular software and virtually nothing for others. While talk may be cheap, it's still a good indicator of popularity.
- **Programming Activity** – some software development is focused into repositories such as GitHub. That allows people to count the number lines of programming code done for each project in a given time period. This is an excellent measure of popularity since writing programs or changing them requires substantial commitment. However, very popular commercial software may not have much user development activity.
- **Popularity Measures** – some sites exist that combine several of the measures discussed here into an overall composite score or rank. In particular, they use programming activity and discussion forums.
- **IT Research Firm Reports** – these firms study the analytics market, interview corporate clients regarding how their needs are being met and/or changing, and write reports describing their take on where each software is now and where they're headed. While I find the reports very interesting reading, they often focus on the company level so it's harder to get package-level information from them.
- **Sales or Download Measures** – the commercial analytics field has undergone a major merger and acquisition phase so that now it is hard to separate out the revenue that comes specifically from analytics. Open source software plays a major role and even the few packages that offer download figures are dicey at best.
- **Competition Use** – organizations that sponsor analytic competitions occasionally report what the winners tend to use. Unfortunately this information is only sporadically available.
- **Growth in Capability** – while *programming activity* (mentioned above) is required before growth in capability can occur, actual growth in capability is a measure of how many new methods of analysis a software package can perform; programming activity can include routine maintenance of existing capability. Unfortunately, most software vendors don't track this measure and, of course, simply counting the number of

new things does not mean they are widely useful new things. I have only been able to collect this data for R, but the results have been very interesting.

## Job Advertisements

One of the best ways to measure the popularity or market share of software for analytics is to count the number of job advertisements for each. Indeed.com is the biggest job site in the U.S. making its sample the best around. As their CEO and co-founder Paul Forster stated, Indeed.com includes "all the jobs from over 1,000 unique sources, comprising the major job boards – Monster, Careerbuilder, Hotjobs, Craigslist – as well as hundreds of newspapers, associations, and company websites." To demonstrate just how dominant its lead is, a search for SPSS (on 2/19/14) showed more than ten times as many jobs on Indeed.com as on its well-known competitor, Monster.com. Indeed.com also has superb search capabilities and it even includes a tool for tracking long-term trends.

Searching for analytics jobs using Indeed.com can be easy, but it can also be very tricky. For many of the analytics software that required only a simple search on its name. However, for software that's hard to locate (e.g. R) or that is general purpose (e.g. Java) it required complex searches and/or some rather tricky calculations which are described here. All of the graphs in this section use those procedures to make the required queries.

Figure 1a shows that Java is in the lead followed by SAS. Python or C, C++/C# are roughly tied for third place. The tie between C and Python is not surprising as many advertisements for analytics jobs that use programming mention both together. (The C variants are combined in a single search since job advertisements usually seek any of them).

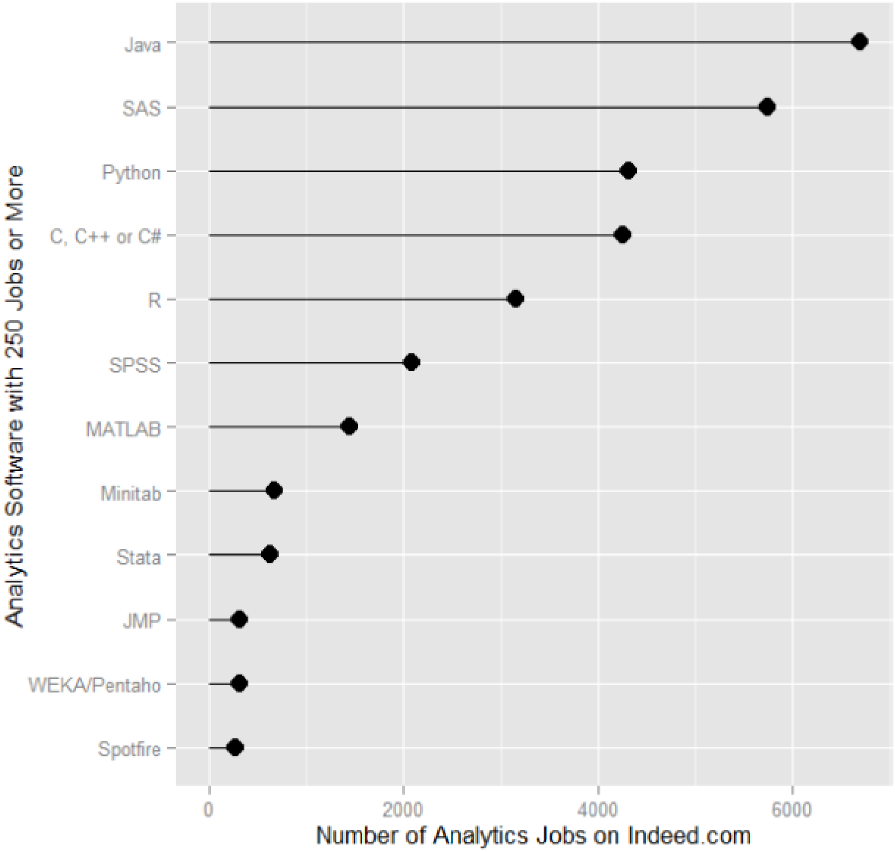


Figure 1a. The number of analytics jobs for the more popular software (250 jobs or more, 2/2014).

R resides in an interestingly large gap between the other domain-specific languages, SAS and SPSS. R has not only caught up with SPSS, but surpassed it with around 50% more job postings. MATLAB has many similarities to R so it's interesting to see that it has only around half the job postings. Note that these are specific to analtyics and MATLAB has many engineering jobs that are not counted in this total.

Much of the software had fewer than 250 jobs. When displayed on the same graph as the industry leaders, their job counts appeared to be zero. Therefore I have plotted them separately in Figure 1b. FICO comes out the leader of this group, followed by Enterprise Miner. Statistica and Alteryx are close to tied at around 55 jobs. From RapidMiner on down, the decline in jobs is fairly smooth. Megaputer's Polyanalyst job count is actually zero.

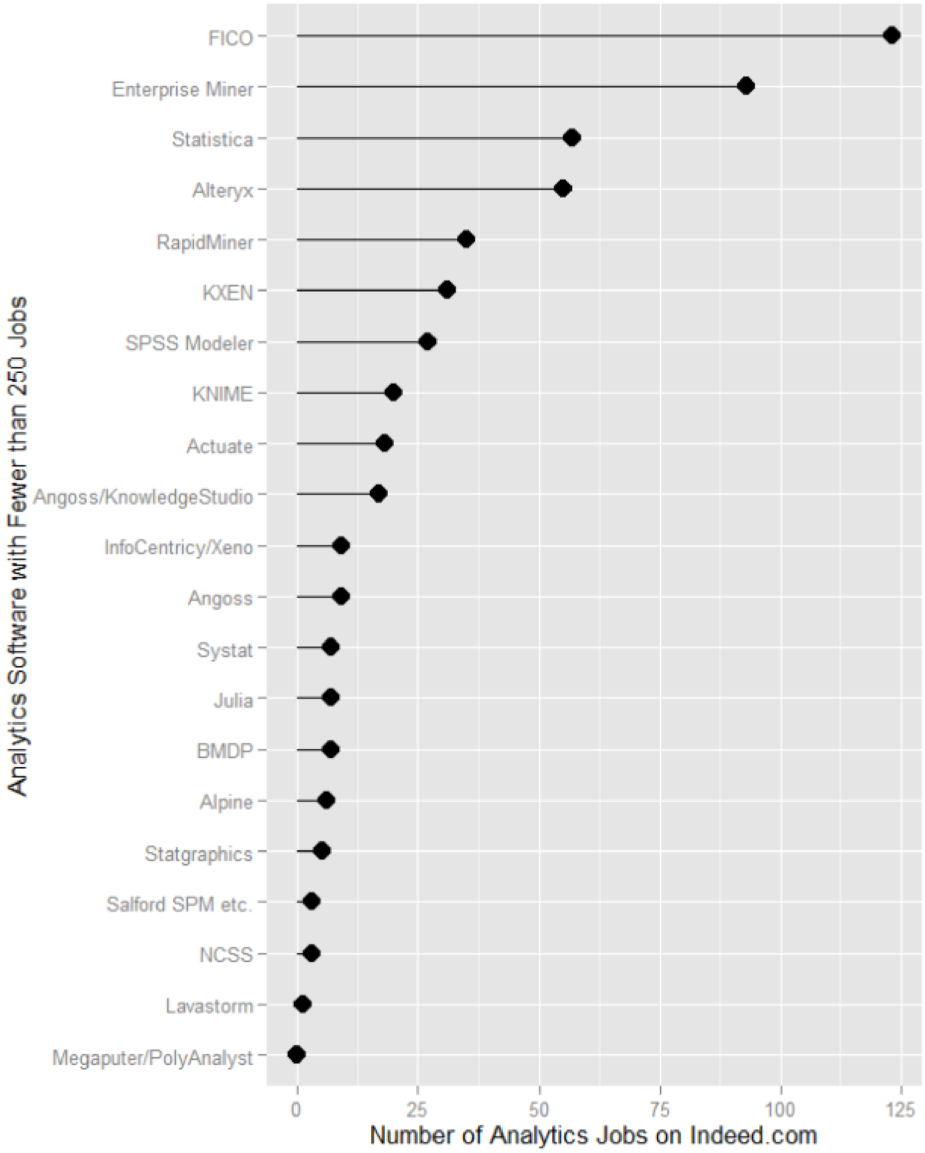


Figure 1b. The number of analytics jobs for the less popular software (under 250 jobs, 2/2014).

It's important to note that the values shown in Figures 1a and 1b are single points in time. The number of jobs for the more popular software do not change much from day to day. Therefore the relative rankings of the software shown in Figure 1a is unlikely to change much over the coming year. The less popular packages shown in Figure 1b have such low job counts that their ranking is likely to shift from month to month. In addition, each software has an overall trend that shows how the demand for jobs changes across the years. You can plot such trends using Indeed.com's Job Trends tool. However, as before, focusing just on analytics jobs requires carefully constructed queries, and when comparing two trends at a time means they *both* have to fit in the same query limit allowed by Indeed.com. Those details are described here.

I'm particularly interested in trends involving R, so let's look at a couple of comparisons. Figure 1c compares the number of analytics jobs available for R and SPSS across time. Analytics jobs for SPSS have not changed much over the years, while those for R have been steadily increasing. The jobs for R finally crossed over and exceeded those for SPSS toward the middle of 2012.

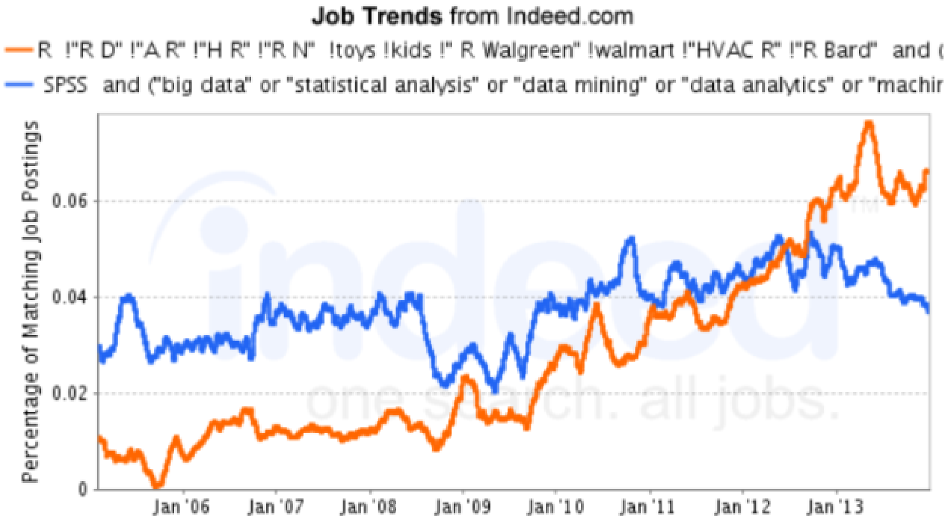


Figure 1c. Analytics job trends for R and SPSS. Note that the legend labels are truncated due to the very long size of the query.

We know from Figure 1a that SAS is still far ahead of R in analytics job postings. How far does R have to go to catch up with SAS? Figure 1d provides one perspective. It would be nice to have the data to forecast when R's growth curve will catch up with SAS's, but Indeed.com does not provide the raw data. However, we can use the approximate slope of each line to get a rough estimate. If jobs for SAS stay level and those for R continue to grow linearly as they have since January 2010, then R will catch up in 3.35 years. If instead the demand for SAS jobs that started in January of 2012 continues, then R will catch up in 1.87 years.

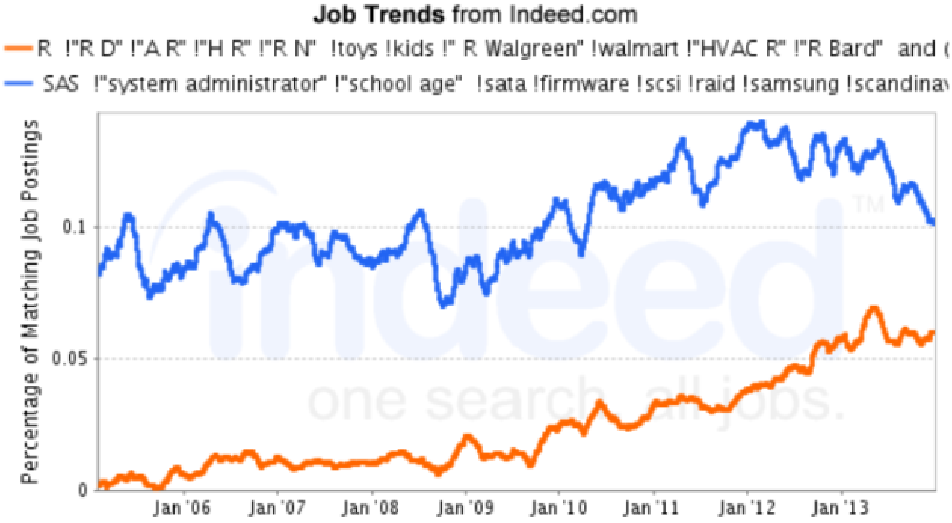


Figure 1d. Analytics job trends for R and SAS. Legend labels are truncated due to long query length.

A debate has been taking place on the Internet regarding the relative place of Python and R. Ironically, this debate about software to do data analytics has involved very little actual data. However it is possible now to at least study the job trends. Figure 1a showed us that Python is well out in front of R, at least on that single day the searches were run. What has the data looked like over time? The answer is in Figure 1e.

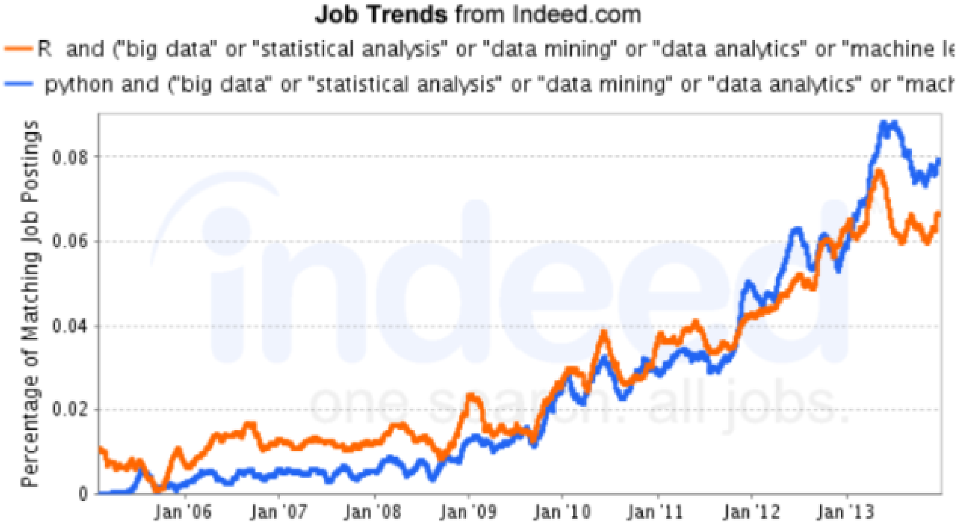


Figure 1e. Jobs trends for R and Python. Legend labels are truncated due to long query length.

Note that in this graph, Python appears to have less of advantage in Figure 1e than it had in Figure 1a. The final point on the trend graph was done only a few days after the queries used in Figure 1a, and that data changed very little in the meantime. The difference is due to the fact that Indeed.com has a limit on query length. Here is the query used for Figure 1e, and the analytic terms it contains were fewer than the one used for Figure 1a.

```
R
and ("big data"
or "statistical analysis"
or "data mining"
or "data analytics"
or "machine learning"
or "quantitative analysis"
or "business analytics"
or "statistical software"
or "predictive modeling")
!"R D" !"A R" !"H R" !"R N"
!toys !kids !" R Walgreen" !walmart
!"HVAC R" !"R Bard"
,
python
and ("big data"
or "statistical analysis"
or "data mining"
or "data analytics"
or "machine learning"
or "quantitative analysis"
or "business analytics"
or "statistical software"
or "predictive modeling")
```

One last trend I considered was for Megaputer's PolyAnalyst. Using the string "Megaputer PolyAnalyst" ("or" is implied) the trend line was completely flat at zero. I only include it here because Gartner considered Megaputer worth including in their Magic Quadrant for Advanced Analytics Platforms report of February 19, 2014.

The detailed description regarding the construction of all the queries used in Figures 1a through 1e is located here.

### Scholarly Articles

The more popular a software package is, the more likely it will appear in scholarly publications as a topic and as a method of analysis. The software that is used in scholarly articles is what the next generation of analysts will graduate knowing, so it's a good leading indicator of where things are headed. Google Scholar offers a way to measure such activity. However, no search of this magnitude is perfect and will include some irrelevant articles and reject some relevant ones. The details of the search terms I used are complex enough to move to a companion article, How to Search For Analytics Articles. Since Google regularly improves its search algorithm, I recollect the data for all years following the protocol described at http://librestats.com/2012/04/12/statistical-software-popularity-on-google-scholar/.

Figure 2a shows the number of articles found for each software package for all the years that Google Scholar can search. SPSS is by far the most dominant package, likely due to its balance between power and ease-of-use. SAS has around half as many, followed by MATLAB and R. Note that the general purpose software MATLAB, Java and Python are included only when found in combination with analytics terms, so view those as much rougher counts than the rest. Neither C nor C++ are included here because it's very difficult to focus the search compared to the search for jobs above, whose job descriptions commonly include a clear target of skills in "C/C++" and "C or C++".

From RapidMiner on down, the counts appear to be zero. That's not the case, but relative to the others, it might as well be.

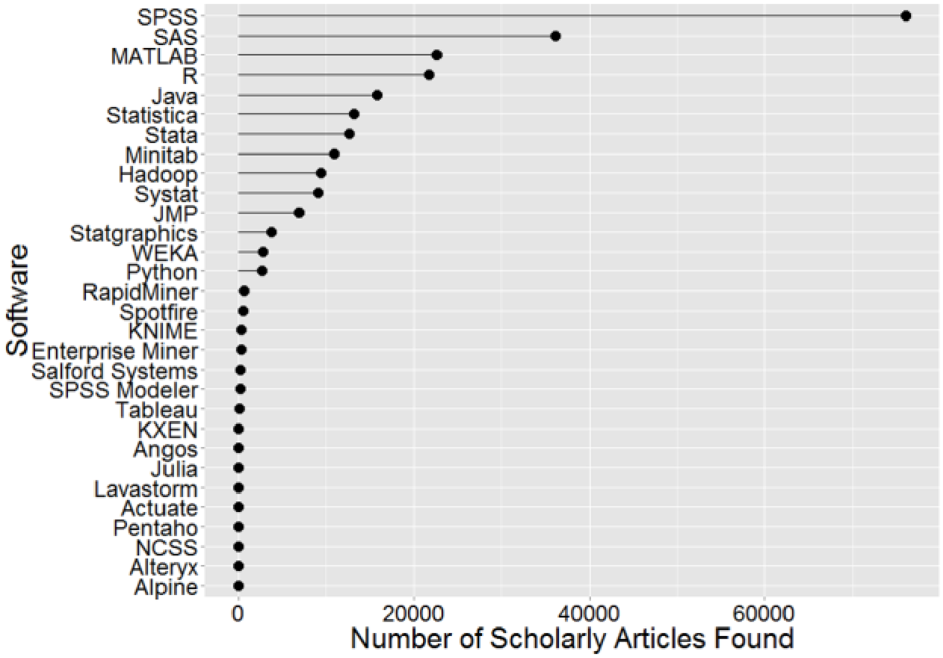


Figure 2a. Number of scholarly articles found for each software.

Figure 2b shows the number of articles for the most popular six classic statistics packages from 1995 through 2013 (the last complete year of data this graph was made). As in Figure 2a, SPSS has a clear lead, but you can see that its dominance peaked in 2007 and its use is now in sharp decline. SAS never came close to SPSS' level of dominance, and it peaked in 2008.

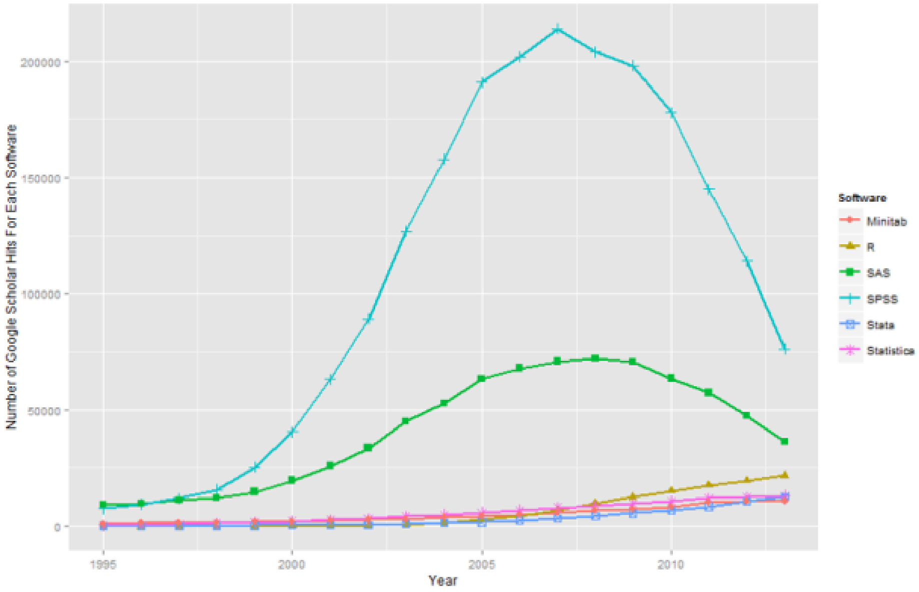


**Figure 2b. Number of scholarly articles found for the top five classic statistics packages.**

Since SAS and SPSS dominate the vertical space in Figure 2a by such a wide margin, I removed those two packages and added the next two most popular statistics packages, Systat and JMP in Figure 2c. Freeing up so much space in the plot now allows us to see that the use of R is experiencing very rapid growth and is pulling away from the pack, solidifying its position in third place. In fact, extending the downward trend of SPSS and the upward trend of R make it likely that sometime during the summer of 2014 R became the most dominant package for analytics used in scholarly publications. Due to the lag caused by the publication process, getting articles online, indexing them, etc. we won't be able to verify that this has happened until well into 2014.

After R, Statistica is in fourth place and growing, but at a much lower rate. Note that in the plots from previous years, Statistica was displayed as a flat line at the very bottom of the graph. That turned out to be a search-related artifact. Many academics who use Statistica don't mention the package by software name but rather say something like, "we used the statistics package by Statsoft."

Extrapolating from the trend lines, it is likely that the use of Stata among academics passed that of Statistica fairly early in 2014. The remaining three packages, Minitab, Systat and JMP are all growing but at a much lower rate than either R or Stata.

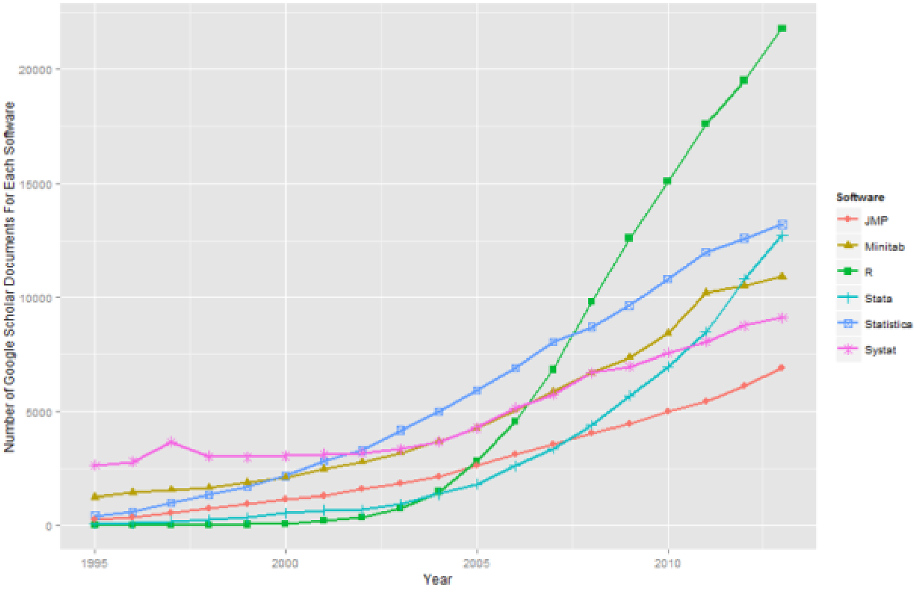


Figure 2c. Number of scholarly articles that reference each software by year, after removing the top two, SPSS and SAS, and adding the next two most popular, Systat and JMP.

## Books

The number of books published on each software reflects their relative popularity. Amazon.com offers an advanced search method which works well for all the software except R. I configured it with the following parameters:

Title: SAS -excerpt -chapter -changes [using SAS as an example]

Subject: Computers & Internet

Condition: New

Format: All formats

Publication Date: After September, 2001 [i.e. 10 years before the search on 10/13/2011]

Since it's difficult to determine how many books use a particular software in its examples, I searched for books that included the software in the *title*. SAS has many manuals for sale as individual chapters or excerpts. Luckily, they contain "chapter" or "excerpt" in their title so I excluded them using the minus sign, e.g. "-excerpt". SAS also has short "changes and enhancements" booklets that the other packages release only in the form of flyers and/or web pages so I excluded "changes" as well.

SAS and SPSS both have many versions of the same book or manual still for sale. For example, Marija Norusis' 3 books on SPSS appear 20 times for various versions of SPSS released in the last 10 years. The SAS and SPSS numbers are both somewhat inflated as a result. Limiting the search to books published in the last 10 years mitigated this problem somewhat, but the SAS and SPSS figures are probably both still somewhat exaggerated.

The count of R books came from http://www.r-project.org/doc/bib/R-books.html. This list does contain seven books on S that are older but still relevant. Version numbers do not appear in any book titles so R avoids the over-counting problem that plagued my count of SAS and SPSS manuals. The most surprising aspect of the result (Figure 3) was how extremely dominant the top few packages are and that three well known packages had no books at all written about them (BMDP, Statistica, Systat). Revolution R and R-PLUS have no books with their names in the titles, but of course the books on R apply to them as well.

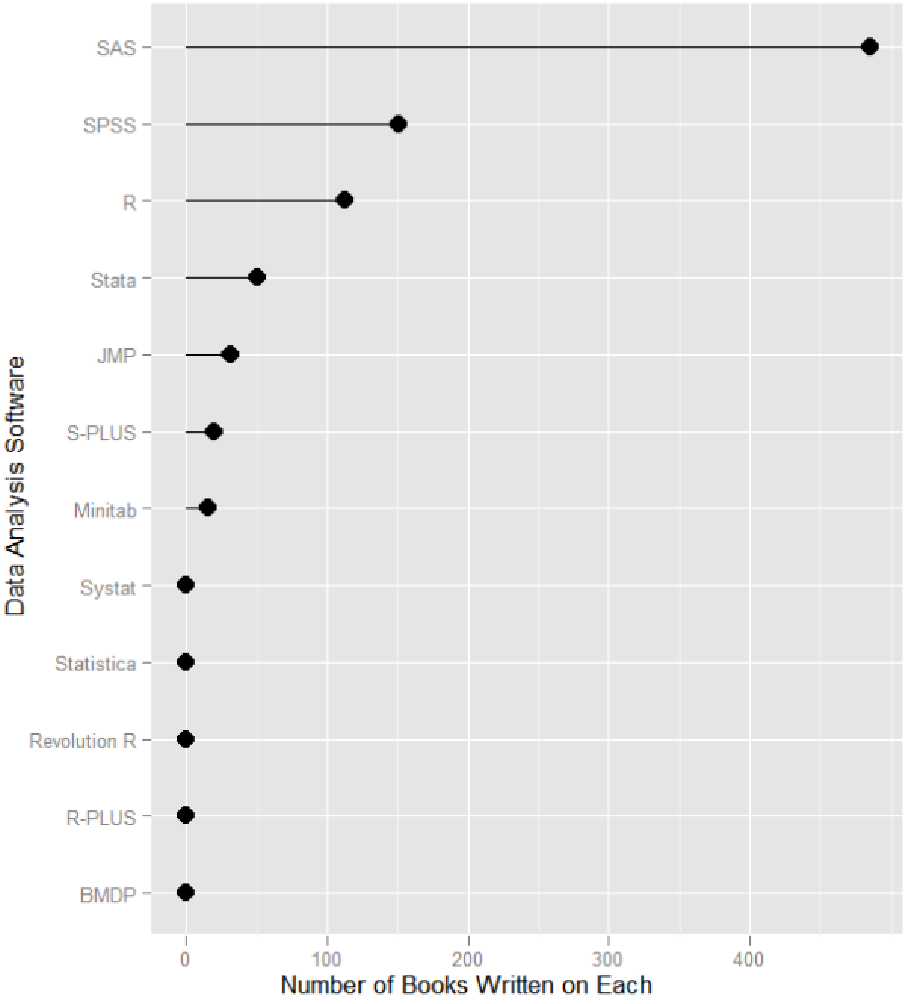


Figure 3. The number of books that contain the name of each software package in their titles.

### Website Popularity

Another measure of software popularity is the number of other web pages that contain links that point to the software's main web site. Figure 4 provides those numbers, recorded using Google on January 5, 2012.

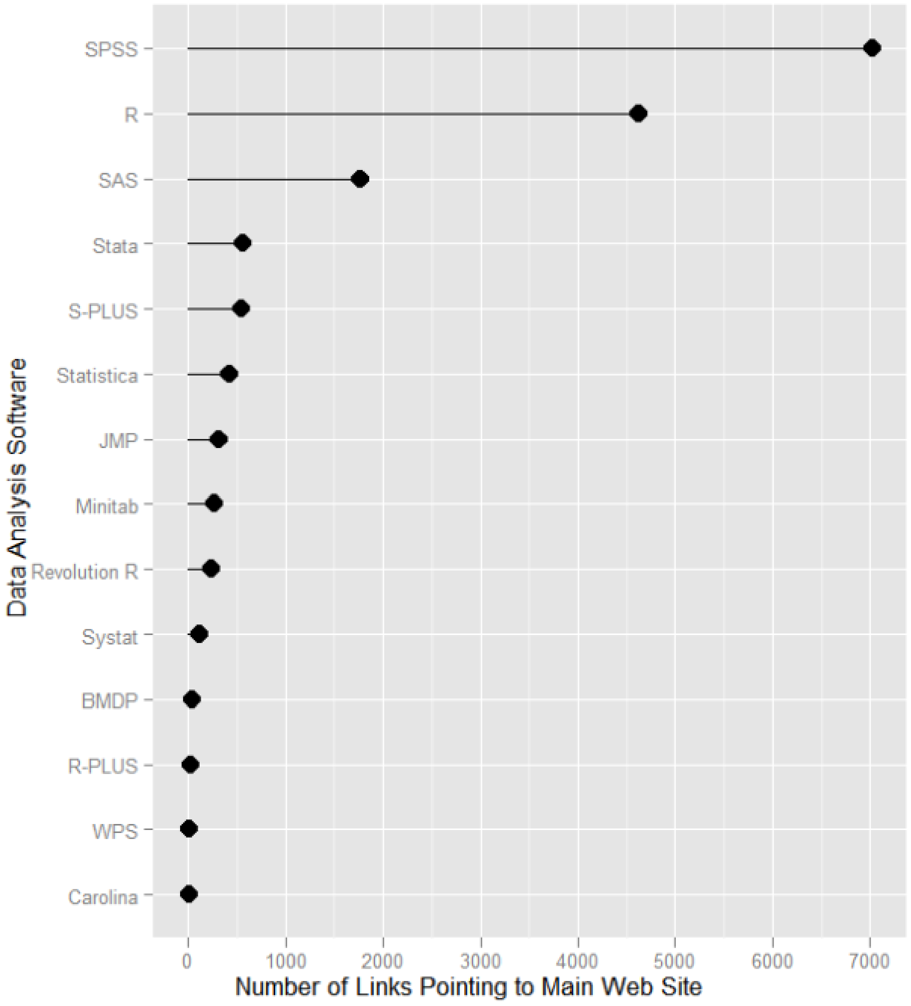


Figure 4. The number of web site links that point to the main web site of each software package.

Now that SPSS is part of IBM, it dominates the results. This reflects the wide range of products that IBM sells, including computer hardware and services that have nothing to do with data analysis. However, the older SPSS.com website no longer shows up early in a web search and the IBM site that it redirects to has a tiny incoming link measure since it is not meant to be a direct link.

R is next in line with a little over half of IBM's measure, followed by SAS with well less than R's value. The other software follows in the order that I suspect is reflective of their respective market shares. Revolution R Enterprise and R-PLUS are commercial versions of R that are relatively quite new to the market. WPS is an implementation of the SAS Language and Carolina is a SAS-to-Java compiler.

The number of incoming links is an important part of Google's famous PageRank algorithm (http://en.wikipedia.org/wiki/PageRank). PageRank is made more useful for searching by (among other things) weighting the importance of each link. Links from major sites like WikiPedia would carry far more weight than would a link from a professor's course syllabus. The practical range of PageRank is from 1 to 10. Figure 9 plots this data (collected on on January 4, 2012). The software appear in tiers, with the two dominant players, SAS and SPSS (IBM), at the highest, and their well-known alternatives one level down. I find it odd that Stata is not in this level. At the very bottom are the World Programming System (WPS) and Carolina, two companies that use the SAS language. There have been quite a few changes in this ranking since last year, with SAS, SPSS and Revolution Analytics moving up one point and R, Stata and Carolina moving down one point. The R-PLUS site maintained its PageRank of 5 this year, which is a bit surprising given that many of its links are broken, and it is in its fourth year of saying, “Be the first to get R-PLUS 3.3″

## Blogs

On Internet blogs, people write about software that interests them, showing how to solve problems and interpreting events in the field. Blog posts contain a great deal of information about their topic, and although it’s not as time consuming as a book to write, maintaining a blog certainly requires effort. Therefore, the number of bloggers writing about analytics software has potential as a measure of popularity or market share. Unfortunately, counting the number of *relevant* blogs is often a difficult task. General purpose software such as Java, Python, the C language variants and MATLAB have many more bloggers writing about general programming topics than just analytics. But separating them out isn’t easy. The name of a blog and the title of its latest post may not give you a clue that it routinely includes articles on analytics.

Another problem arises from the fact that what some companies would write up as a newsletter, others would do as a set of blogs, where several people in the company each contribute their own blog, but they’re also combined into a single company blog. Statsoft and Minitab offer examples of this. So what’s really interesting is not company employees who are assigned to write blogs, but rather those written by outside volunteers. In a few lucky cases, lists of such blogs are maintained, usually by blog consolidators, who combine many blogs into a large “metablog.” All I have to do is find such lists and count the blogs. I don’t attempt to extract the few vendor employees that I know are blended into such lists. I only skip those lists that are exclusively employee-based (or very close to it). The results are shown in Table 1.

| | Number | |
|---|---|---|
| **Software** | **of Blogs** | **Source** |
| R | 550 | R-Bloggers.com |
| Python | 60 | SciPy.org |
| SAS | 40 | PROC-X.com, sasCommunity.org Planet |
| Stata | 11 | Stata-Bloggers.com |

Table 1. Number of blogs devoted to each software package on April 7, 2014, and the source of the data.

R’s 550 blogs is quite an impressive number. For Python, I could only find that list of 60 that were devoted to the SciPy subroutine library. Some of those are likely cover topics besides analytics, but to determine which never cover the topic would be quite time consuming. The 40 blogs about SAS is still an impressive figure given that Stata was the only other company that even garnered a list anywhere. That list is at the vendor itself, Statacorp, but it consists of non-employees except for one.

While searching for lists of blogs on other software, I did find individual blogs that at least occasionally covered a particular topic. However, keeping this list up to date is far too time consuming given the relative ease with which other popularity measures are collected.

If you know of other lists of relevant blogs, please let me know and I'll add them. If you're a software vendor employee reading this, and your company does not build a metablog or at least maintain a list of your bloggers, I recommend taking advantage of this important source of free publicity.

**Discussion Forum Activity**

There are some stable and objective measures regarding analytic software. Schwartz (2009) suggested estimating relative popularity by plotting the amount of email discussion devoted to each. The most widely used packages all have discussion lists, or "listservs" devoted to them. The less popular ones either do not have such discussions or, like the lists for Minitab or S-PLUS, may have only a dozen or so emails per year. Some software packages have multiple discussion lists. For example, there are 25 devoted to using R (http://www.r-project.org/mail.html). Topics range from general help to various focused areas such as  graphics, mapping, ecology, epidemiology, etc. . A broader list, including a version of R-Help in Spanish, lists 48 discussions (https://stat.ethz.ch/mailman/listinfo).

Figure 1a shows the level of activity on only each main discussion listserv in a typical month (i.e. forums, news groups and Google groups are excluded). Each point represents the sum of the 12 monthly counts that occurred in that year. This plot contains data through the end of 2012. If you read this article in previous years, this plot used to display the mean number of emails per month rather than the sum. Therefore the scale of the *y*-axis is different but the relative locations of the points are virtually identical. I made this change to enable better a better comparison to discussion forums (e.g. Fig. 5a).

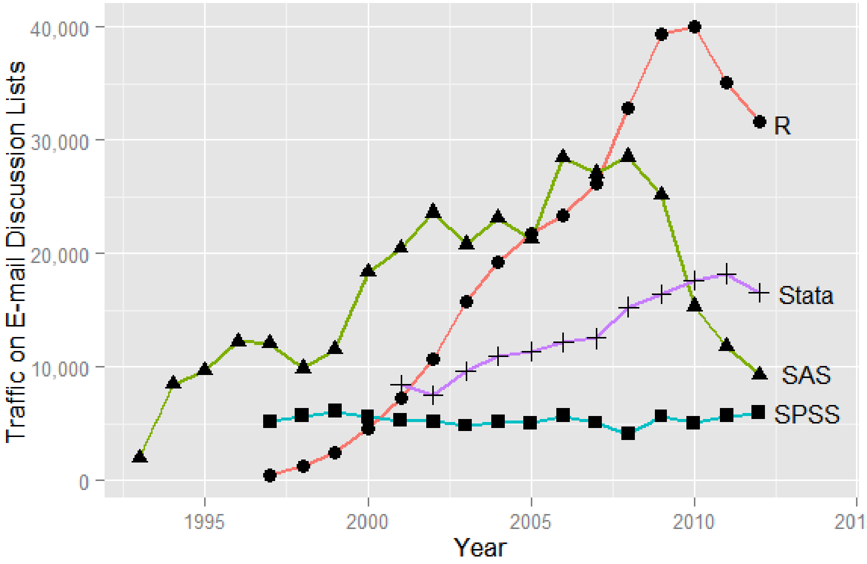


Figure 5a. Sum of monthly email traffic on each software's main listserv discussion list.

We can see that discussion of R has grown the most rapidly and, for the past few years, R is the most discussed software by an almost two-to-one margin. In recent years, it is followed by Stata, SAS and SPSS, respectively.

Stata showed steady discussion growth until it passed SAS in 2010.

SAS saw rapid growth in its discussion until 2006 when it leveled off and then declined. That decline coincided with the strong growth of both R and Stata, offering competition to SAS.

SPSS held steady at a low rate across the time frame, which may be attributable to its great ease of use relative to the other packages. With both the interface and the documentation aimed at people who prefer GUIs over programming, there's less need to ask how to do variations on an analysis. In fact, there's less *ability* to do such variations. As a result, I doubt SPSS' low showing in this graph is indicative of its popularity or market share.

It would be interesting to see what topics were most discussed on each list. The only such analysis of which I am aware was done by Arthur Tabachnek (2010) for the SAS list. The most popular topic in 2009 turned out to be…R! You can read his full analysis here under *slides from the 2010 session*.

From 2011 onward, R and Stata joined SAS in the decline in listserv discussion. Given the sharp increase in the popularity of business analytics, Big Data, and so on, it is unlikely that people are using or talking about these tools less. Instead, alternative forums of discussion have appeared. The site Stack Overflow (http://stackoverflow.com) covers a wide range of programming and statistical topics, while its sister site, Cross Validated (http://stats.stackexchange.com/), focuses only on statistical analysis. A third site, Talk Stats (http://www.talkstats.com), also focuses on statistical analysis. At all three sites, users tag their topics making it particularly easy to focus searches. Figure 5b shows the software people are discussing there.

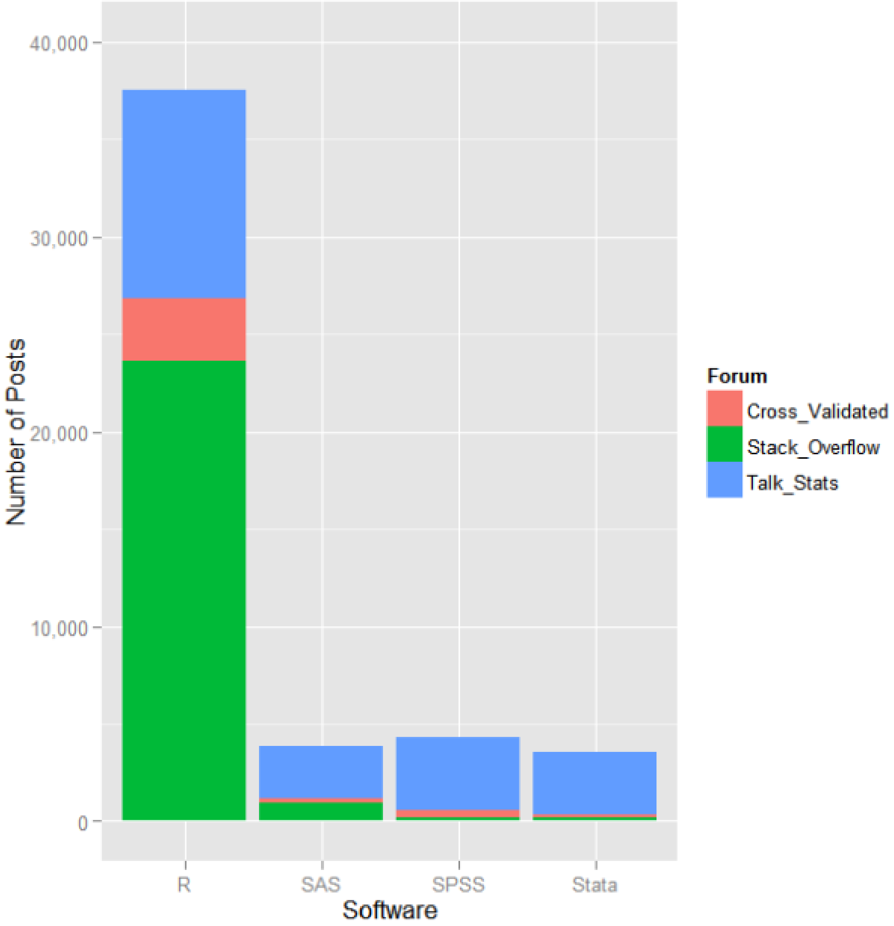


Figure 5b. Number of posts per software on each forum on 2/10/2013.

We can see that the discussion of R is dramatically higher than the other packages, which don't differ very much among themselves. Much of this difference is due to the influence of Stack Overflow, reflecting the vastly greater popularity of R as a programming language. However, even removing that effect, it is easy to see that R still dominates the discussions on the more statistically-oriented forums. This data is cumulative, but we can get a yearly view of just two of the tags: R and SAS (Figure 5c).

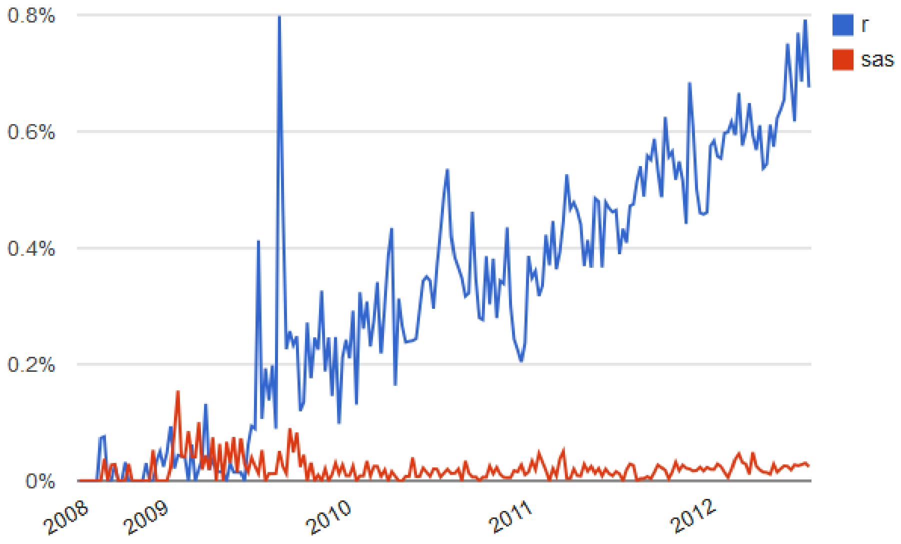


Figure 5c. Number of R- or SAS-related posts to Stack Overflow by week.

We see that discussion of SAS and R were roughly comparable until mid-2009 when the discussion of R began its very rapid climb. The page that provides this data does not display data for SPSS or Stata. The amount of data may be too low; no message provides the reason (see http://hewgill.com/~greg/stackoverflow/stack_overflow/tags).

Other popular discussion forum sites are LinkedIn.com and Quora.com. Neither of these sites make it easy to count number of posts, but they do display the number of people who have joined discussion groups (Figure 5d).

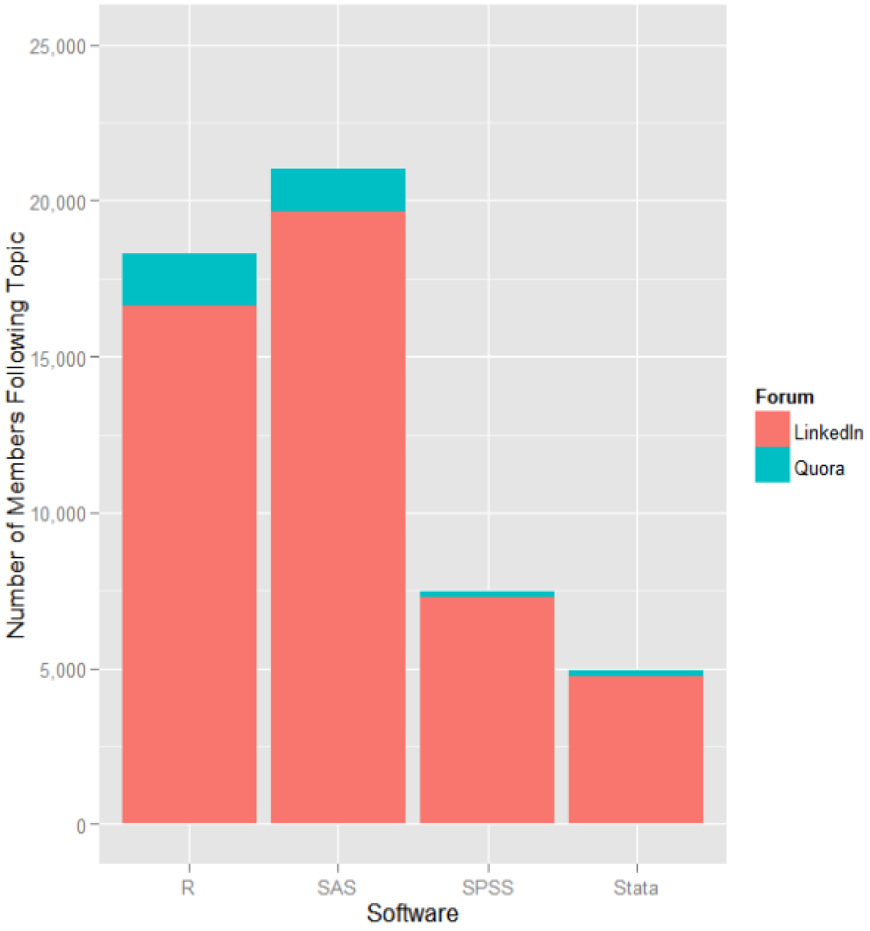


Figure 5d. Number of people registered in the main discussion group for each software on 2/10/2013.

In Figure 1d we get a better view of corporate software use. I do not know the ratio of corporate to academic use of LinkedIn, but among the academics I do know (quite a few) they use it very little. In this world, SAS is the leader with R close behind. It's interesting to see SPSS with a 50% lead over Stata; it was also slightly higher in Fig. 1b. Remember these are people who have joined a group, not necessary people who are talking as the previous two figures were. Still, group membership should be a reasonable proxy for popularity or market share.

### Programming Activity

This section is planned for future expansion. Stay tuned.

### Popularity Measures

The TIOBE Community Programming Index ranks the popularity of programming languages, but from a programming language perspective rather than as analytical software (http://www.tiobe.com). It extracts measurements from blogs, entries in Wikipedia, books on Amazon, and search engine results, and combines them into a single index.  In January 2012, they ranked R in 24th place and SAS at 31st. However, by February 2014, the two had reversed positions with SAS in 21st place and R in 44th.

The only other language that focuses on data analysis that is ranked in the top 100 are S and S-PLUS (R is an implementation of the S language, as is S-PLUS). In previous years SPSS ranked in the 50-100 group but by

February of 2013 it had dropped out (and is still out in February 2014.)

The Transparent Language Popularity Index is very similar to the TIOBE Index with except that its ranking software, algorithm and data are published for all to see. Their latest figures on 2/15/2014 were from July of 2013, at which time it ranked R in 14th place and SAS in 31st. This index also ranks R as a scripting language, where it is in 6th place after tools like PHP, Python and Perl. SAS is also ranked 5th in the "Other" category, when compared to languages such as like COBOL or PL/SQL. While these two additional areas may seen irrelevant to data analysis, it's good to know both these tools have more flexibility than most other domain specific languages which focus on data analysis.

Langpop.com also ranks programming languages (http://langpop.com/) in a variety of interesting ways, but unfortunately their focus excludes statistical software.

**Surveys of Use**

One way to estimate the relative popularity of data analysis software is though a survey. Rexer Analytics does a survey every other year asking a wide range of questions regarding data mining. Figure 6a shows the results of the question about the tools that respondents reported using in 2013.

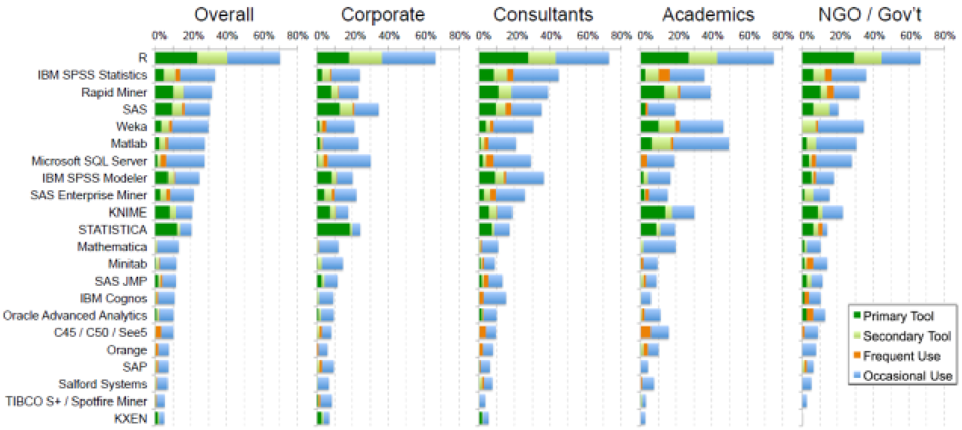


Figure 6a. Rexer Analytics Data Miner Survey 2013.

We see that R comes out on top by a wide margin, with 70% of data miners using it. SPSS, RapidMiner, SAS and Weka all follow with only around 30% of users. The entire report contained over 40 questions on topics such as algorithms used, fields, challenges, data, impact of the economy on the field, and more. It's interesting to note that while this survey is aimed at data miners, SPSS and SAS are used more often than their more expensive products aimed specifically at data mining, IBM SPSS Modeler and SAS Enterprise Miner.

The results of a similar poll done by the data mining web site KDnuggets in 2013 are shown in Figure 6b. This one shows RapidMiner in first place with 39.2% of users reporting having used it for a real project. R follows closely behind with 37.4% of users. There's quite a large gap in which Excel resides, with 28% of users. Weka/Pentaho and Python are tied at 14.3%, followed by the rest. Note that both RapidMiner and KNIME are listed twice, once in their free version and again in commercial. While it's tempting to add these two, there may be overlap for organizations that use both.

It's interesting to note that four of the top five packages used were open source.

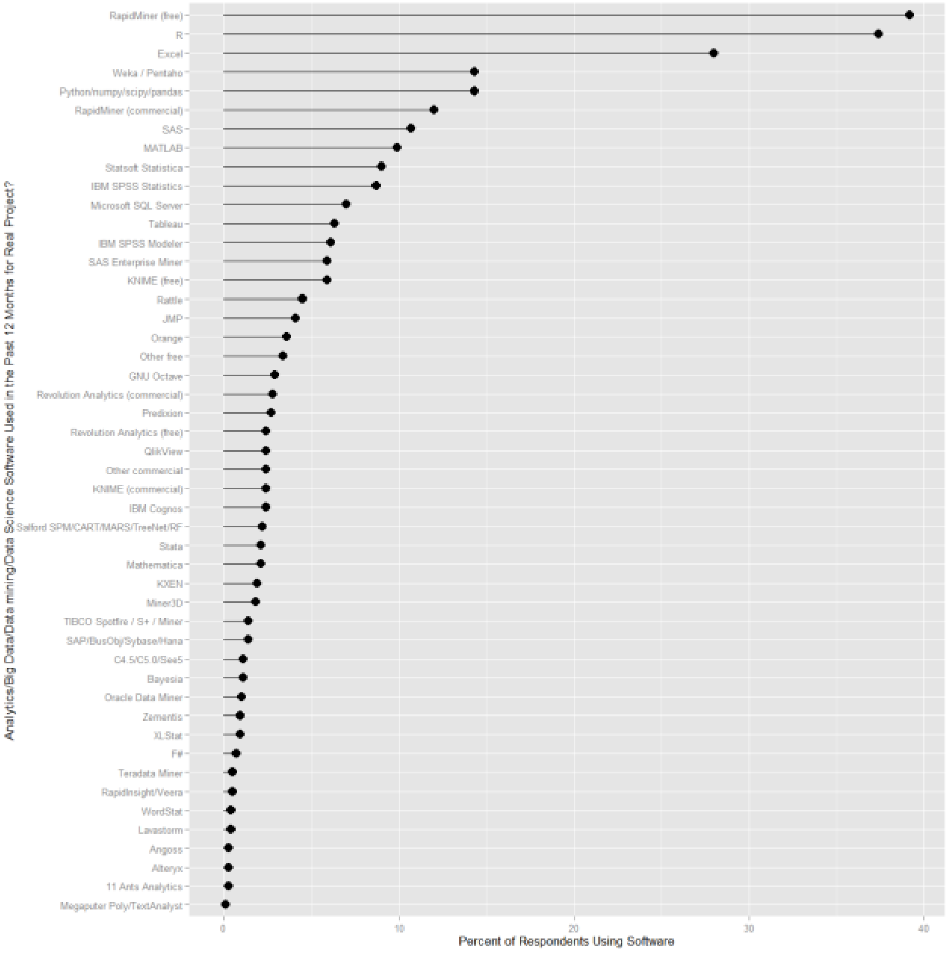


Figure 6b. Percent of respondents that used each software in KDnuggets 2013 poll.

The KDnuggets site conducted similar poll, this time asking, “What programming languages you used for data mining / data analysis in the past 12 months?”  R dominated this poll with over 60% of respondents, as shown in Figure 6c. Python and SQL followed with around 37% each.

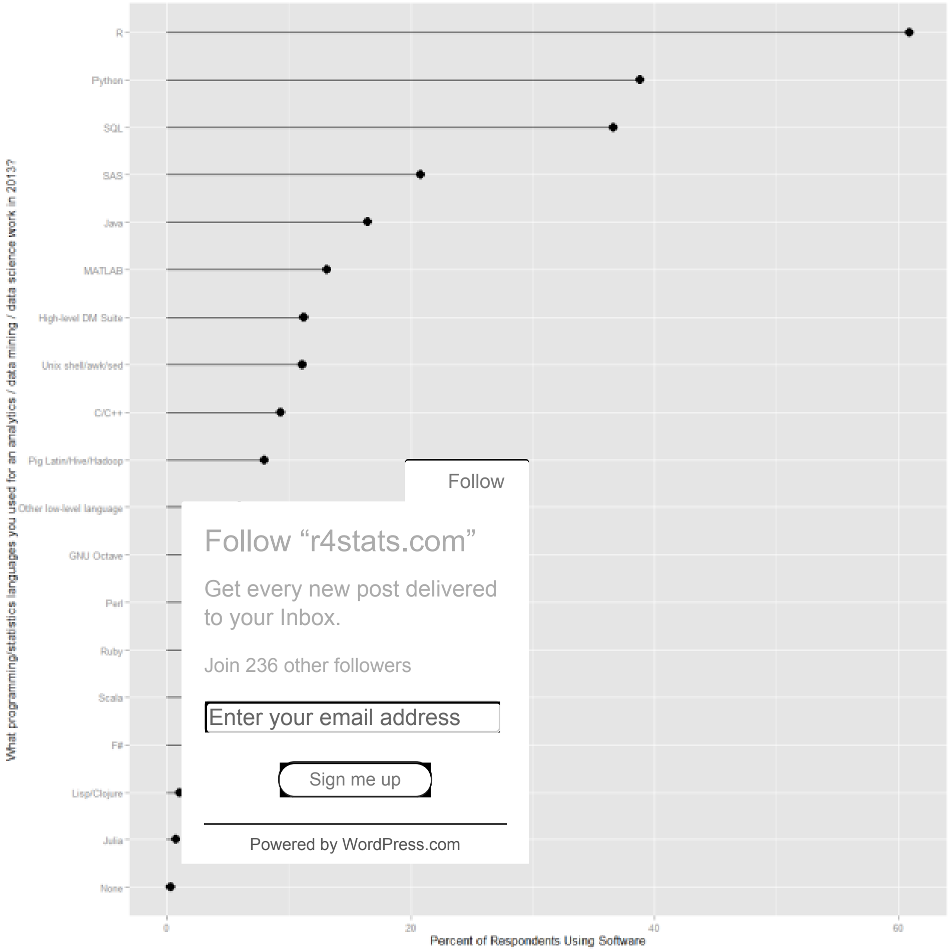


Figure 6c. KDnuggets poll on programming tools used for an analytics / data mining / data science work in 2013.

O’Reilly Media conducted a survey of conference attendees in 2012 and 2013 of the Strata Conference: Making Data Work and Strata + Hadoop World. The top bar in Figure 6d shows that 57% of respondents listed some form of data analysis as their primary job. SQL is listed as the top tool with 71% of respondents using it. This indicates that most attendees stored their data in relational databases. R came out as the top tool for advanced analytics, with 43% of respondents using it. Given that some respondents were attending Hadoop world, it’s not surprising that both Hadoop and the related Mahout came out higher here than in most other measures of popularity discussed in this paper. The fact that SAS/SPSS came out the bottom is another clear indication that this is not a random sample of analytics users.

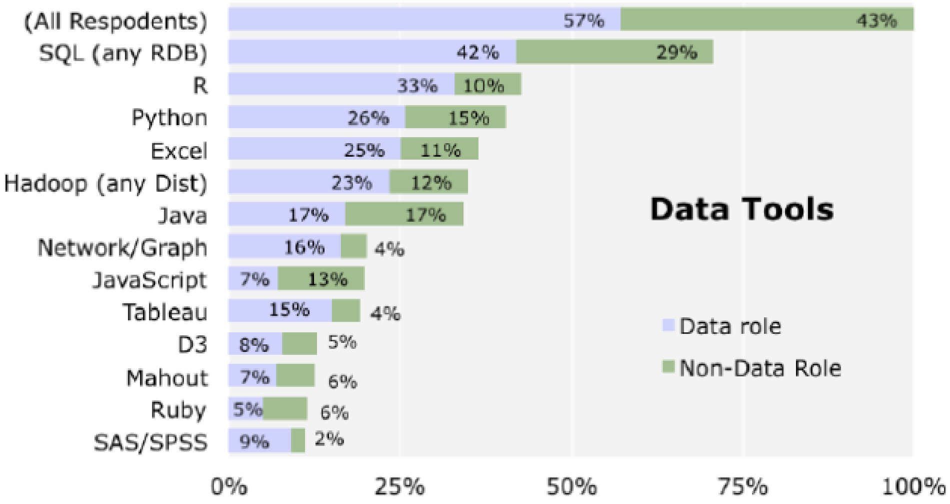


Figure 6d. O'Reilly Data Science Survey results for 2012 and 2013 combined.

Lavastorm, Inc. conducted a survey of analytic communities including LinkedIn's Lavastorm Analytics Community Group, Data Science Central and KDnuggets. The results were published in March, 2013, and the bar chart of "self-service analytic tool" usage among their respondents is shown in Figure 6e. Excel comes out as the top tool, with 75.6% of respondents reporting its use. While other surveys show Excel use similarly high, some don't include it at all, leaving us to wonder what survey researchers and respondents were thinking regarding this relatively low-powered tool.

R comes out as the top advanced analytics tool with 35.3% of respondents, followed closely by SAS. MS Access' position in 4th place is a bit of an outlier as no other surveys include it at all. Lavastorm comes out with a much higher market share (3.4%) than the KDNuggests poll indicated (0.4%), but that's hardly a surprise given than the survey was aimed at the Lavastorm's LinkedIn community group.

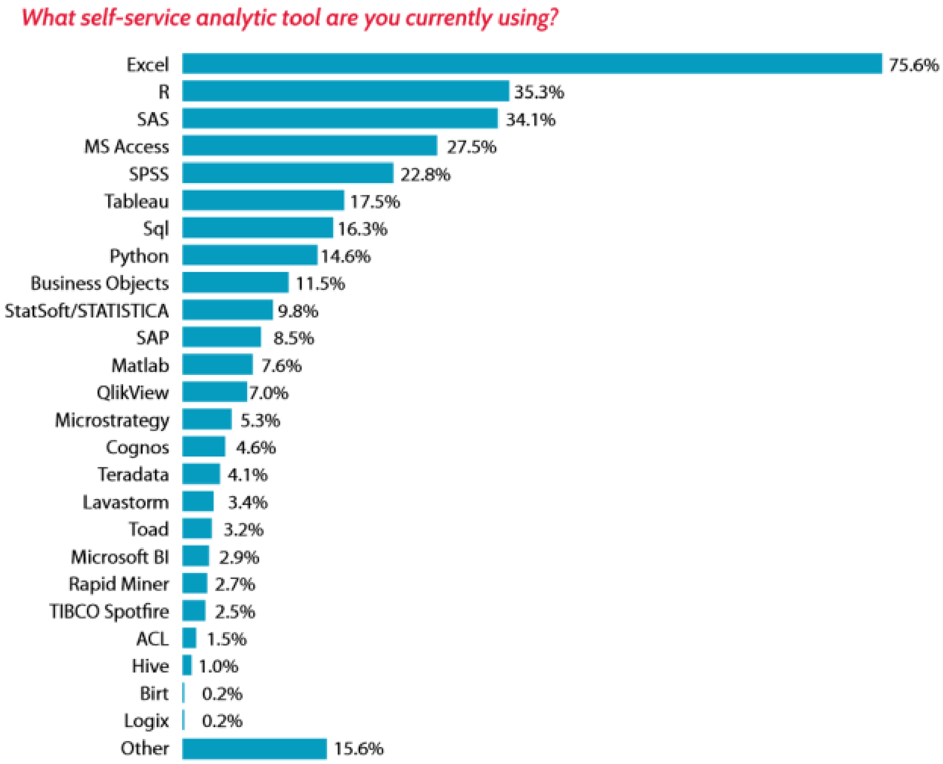


Figure 6e. Lavastorm survey of analytics tools.

## IT Research Firms

IT research firms study software products and corporate strategies and provide their opinions on each in reports they sell to their clients. Each company has its own criteria for rating companies, so they don't always agree, but I find the reports extremely interesting reading.

Gartner, Inc. is one of the companies that provides such reports.  The "Magic Quadrant" from their report, *Advanced Analytics for Business Analysts* is shown in Figure 7a. Since it rates companies, strictly open source software such as R, Python and Java are not shown. IBM (SPSS) and SAS are the leading companies, which comes as no surprise. I was, however, surprised by the inclusion of RapidMiner and KNIME in the Leaders' quadrant. Both products are available in free, open source versions and in commercial versions. Both have done also very well in user surveys (see *Surveys of Use* section below). However, user surveys and analyst reports often disagree.

Another surprise in this figure is the inclusion of Megaputer (PolyAnalyst). There were actually zero jobs for that product in Figure 1b. But they're shown as being in the same neighborhood as Oracle and Microsoft!

Revolution Analytics, a company whose main business is providing a commercial version of R, is on the edge of the Leaders' quadrant. The full report provides an insightful analysis of the strengths and weaknesses of each company's offerings.

Thanks to Alteryx, Inc. the 2014 Gartner Group report on Advanced Analytics is available here. Note that the links that provide such reports for free tend to expire often but you can usually find each by searching the internet for the report names.

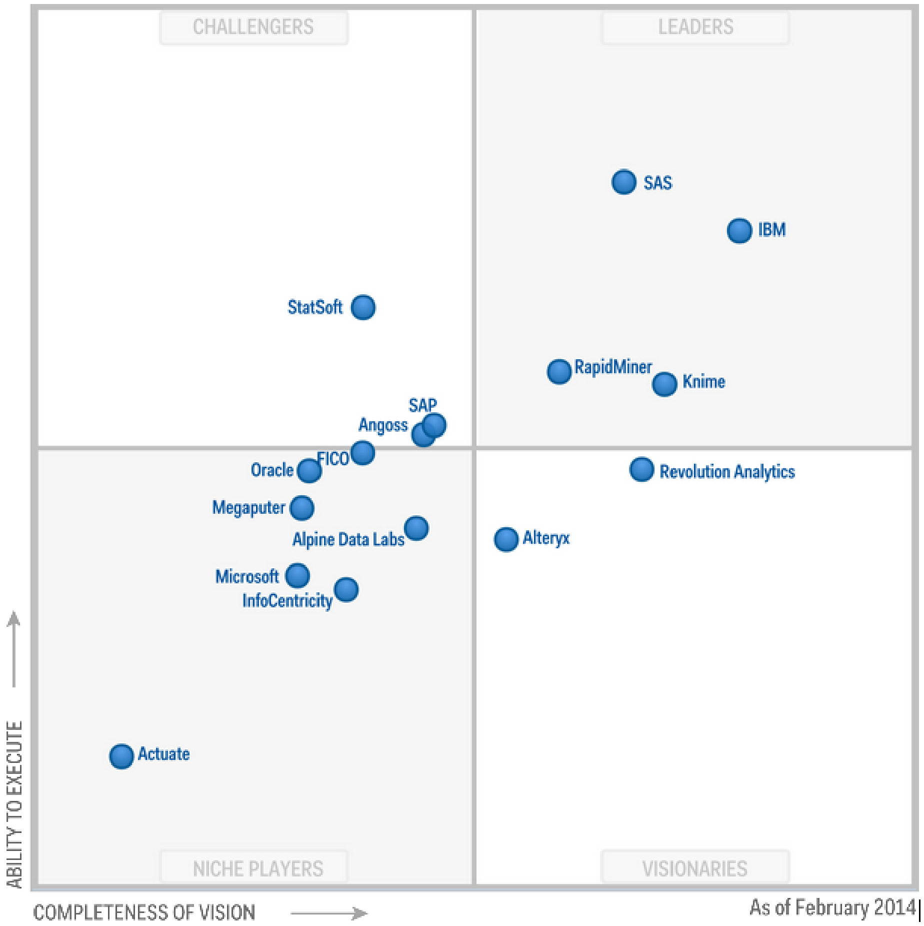


Figure 7a. Gartner "Magic Quadrant" plot of companies that sell advanced analtyics software (2014).

Forrester Research, Inc. is another company that provides similar reports. It's "Wave" plot from their report, "The Forrester Wave: Big Data Predictive Analytics Solutions, Q1 2013″ is shown in Figure 7b.

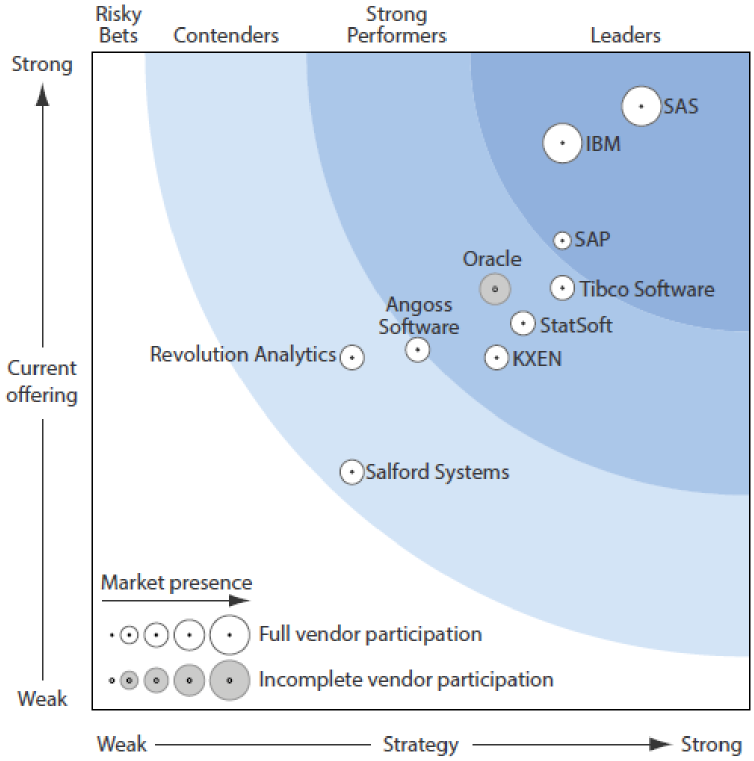


Figure 7b. Forrester Wave plot of big data predictive analytics solutions, Q1 2013.

Again, IBM (SPSS) and SAS are the strongest companies, but that seems to be all that the two reports seem to agree on! The reports emphasize different aspects of the companies being rated, which accounts for the radically different plots. You can read the full Forrester report, compliments of SAP, here.

### Sales & Downloads

Sales figures reported by some commercial vendors include products that have little to do with analysis. Many vendors don't release sales figures, or they release them in a form that combines many different products, making the examination of a particular product impossible. For open source software such as R (Ihaka and Gentleman 1996) you could count downloads, but one confused person can download many copies, inflating the total. Conversely, many people can use a single download on a server, deflating it.

Download counts for the R-based Bioconductor project are located at http://www.bioconductor.org/packages/stats/. Similar figures for downloads of Stata add-ons (not Stata itself) are available at http://fmwww.bc.edu/fmrc/reports/Report.SSC.html. A list of Stata repositories is available at http://stata.com/links/resources2.html. The many sources of downloads both in repositories and individuals' web sites makes counting downloads a very difficult task.

### Competition Use

Kaggle.com is a web site that sponsors data analysis contests. People post data analysis problems there along the amount of money they are willing pay the person or team who solves their problem the best. Figure 8 shows the software used by the data analysts working on the problems. R is in the lead by a wide margin. R's dominance is even greater among the contest winners, over 50% of whom used R. A potential source of bias in these figures is that the licenses of most proprietary software prohibits its use for the benefit of outside organizations (universities can help federal grant-providing agencies such as NSF and NIH, but cannot even solve problems for government agencies in general or nonprofits). However, I manage the research software site licenses at the University of Tennessee, and I can attest to the fact that people are often unaware of this limitation. (Note that as of 4/11/2014 this graph of 2011 data is still Kaggle's most current graph.)

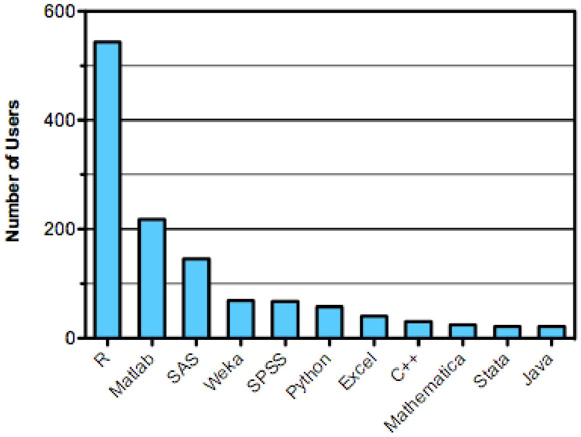


Figure 8. Software used in data analysis competitions in 2011.

### Growth in Capability

The capability of analytics software has grown significantly over the years. It would be helpful to be able to plot the growth of each software package's capabilities, but such data are hard to obtain. John Fox (2009) acquired them for R's main distribution site http://cran.r-project.org/. I collected the data for later versions following his method.

Figure 9 shows that the growth in R packages is following a rapid parabolic arc (quadratic fit with R-squared=.998). The right-most point is for version 3.0.2, the last version released in 2013.

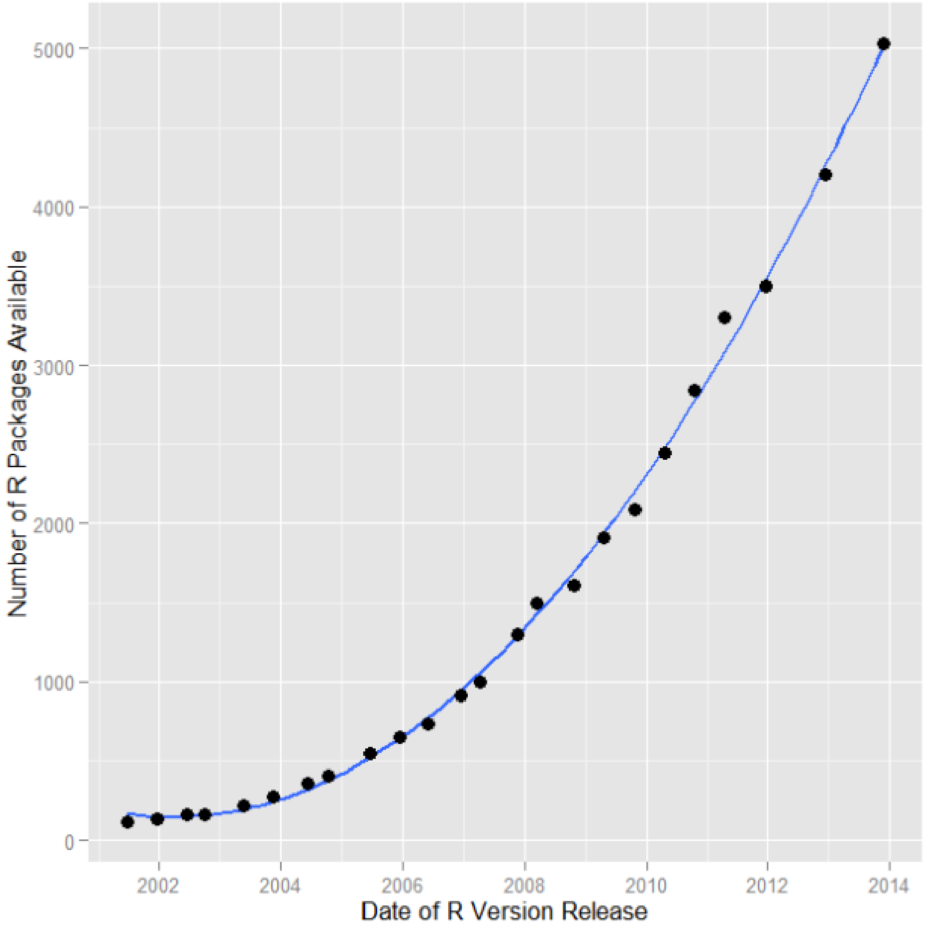


Figure 9. Number of R packages plotted for each major release of R.

To put this astonishing growth in perspective, let us compare it to the most dominant commercial package, SAS. In version, 9.3, SAS contains around 1,200 commands that are roughly equivalent to R functions (procs, functions etc. in Base, Stat, ETS, HP Forecasting, Graph, IML, Macro, OR, QC). In 2013, R added 835 packages, counting only CRAN, or approximately 17,390 functions. *During 2013 alone, R added more functions/procs than SAS Institute has written in its entire history!*

Of course SAS and R commands are not perfectly equivalent. Some SAS procedures have many more options to control their output than R functions do. However, R functions can nest inside one another, creating nearly infinite combinations. Also, SAS is now out with version 9.4 and I have not repeated the arduous task of recounting its commands. If SAS Institute would provide the figure, I would be happy to list it here. While the comparison is not perfect, it does provide an interesting perspective on the size and growth rate of R.

As rapid as R's growth has been, these data represent only the main CRAN repository. R has eight other software repositories, such as Bioconductor, that are not included in
Figure 8. A program run on 4/11/2014 counted 7,380 R packages at all major repositories, 5,339 of which were at CRAN. So the growth curve for the software at all repositories would be roughly 38% higher on the y-axis than the one shown in Figure 8. The total growth in R functions for 2013 was approximately 17,390 * 1.38 or 23,998.

As with any analysis software, individuals also maintain their own separate collections typically available on their web sites. Those are not easily counted.

What's the total number of R functions? The Rdocumentation site shows the latest figures counts of both packages and functions on CRAN. They indicate that there are an average of 20.826 functions per package. Since a program on 4/7/2014 counted 7,364 R packages at all major repositories, on that date there were approximately 153,696 total functions in R, over an order of magnitude more than commands in SAS.

## What's Missing?

I previously included on Google Trends. That site tracks not what's actually on the Internet via searches, but rather the keywords and phrases that people are entering into their Google searches. That ended up being so variable as to be essentially worthless. For an interesting discussion of this topic, see this article by Rick Wicklin.

### Conclusion

[This section is needs an overhaul due to the new software added 2/20/14.] I'm interested in other ways to measure software popularity. If you have any ideas on the subject, please contact me at muenchen.bob@gmail.com.

If you are a SAS or SPSS user interested in learning more about R, you might consider my book, *R for SAS and SPSS Users*. Stata users might want to consider reading *R for Stata Users*, which I wrote with Stata guru Joe Hilbe. I also teach workshops quarterly on these topics with Revolution Analytics.

### Acknowledgments

I am grateful to the following people for their suggestions that improved this article: John Fox (2009) provided the data on R package growth; Marc Schwartz (2009) suggested plotting the amount of activity on e-mail discussion lists; Duncan Murdoch clarified the pitfalls of counting downloads; Martin Weiss pointed out both how to query Statlist for its number of subscribers; Christopher Baum provided information regarding counting Stata downloads; John (Jiangtang) HU suggeseted I add more detail from the TIOBE index; Andre Wielki suggested the addition of SAS Institute's support forums; Kjetil Halvorsen provided the location of the expanded list of Internet R discussions; Dario Solari and Joris Meys suggested how to improve Google Insight searches; Keo Ormsby provded useful suggestions regarding Google Scholar; Karl Rexer provided his data mining survey data; Gregory Piatetsky-Shapiro provided his KDnuggets data mining poll; Tal Galili provided advice on blogs and consolidation, as well as Stack Exchange and Stack Overflow; Patrick Burns provided general advice; Nick Cox clarified the role of Stata's software repositories and of popularity itself; Stas Kolenikov provided the link of known Stata repositories; Rick Wicklin convinced me to stop trying to get anything useful out of Google Insights; Drew Schmidt automated the collection of the data in Figures 7a and 7b; Francois Briatte provided the link that creates Figure 1c; Rasmus Bååth provided the median number of functions in an R package.

### Bibliography

J. Fox. Aspects of the Social Organization and Trajectory of the R Project. *R Journal*, http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Fox.pdf

R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

R. Muenchen, *R for SAS and SPSS Users*, Springer, 2009

R. Muenchen, J. Hilbe, *R for Stata Users*, Springer, 2010

M. Schwartz, 1/7/2009, http://tolstoy.newcastle.edu.au/R/e6/help/09/01/0517.html

**Trademarks**

BMDP, Carolina, JMP, Minitab, R-PLUS, Revolution R, SAS, SAS Enterprinse Miner, IBM SPSS Modeler, IBM SPSS Statistics, Stata, Statistica, Systat and WPS are registered trademarks of their respective companies.

**Share this:**

Email | Twitter 206 | Facebook 483 | Google

Like

5 bloggers like this.

## 61 Responses to *The Popularity of Data Analysis Software*

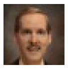
**bob mcconnaughey** *says:*

June 1, 2012 at 2:07 pm

i'm not surprised that R, in particular, has done spectacularly well with respect to analytic use – it has, as best i can tell, virtually all the analytic tools one might need. I've been worried for decades now about the ever increasing use of excel for both data mgt and analysis. So many projects/"data sets"/ analyses have come our way in excel spreadsheets only for major problems in data integrity, tracing flow of data changes that led to errors, even analyses that were later found to be completely hosed because a user had done something as simple, and as deadly, as sorting a column instead of the records/rows.

What SAS has and,really should concentrate on, is its data handling, manipulation, organization, data validation features..that are all built into Base SAS. I have, and appreciate, your R for SAS/SpSS Users – and i can't help but think that organizations that rely on both "data integrity" which, really, is SAS' great strength and analysis could profitably use SAS for complex data manipulations and then write out files in one of the many formats R takes, do the analytics in R and pull the results back into Base SAS. A few months ago i helped out a friend who was analyzing generational data drawn from 80 + yrs from the complete medical birth registry of Norway. SPSS is the data manipulation software they use..and the task of linking families, sibs, half sibs with flags/subsets for individuals/families that had various birth defects over multiple generations was seemingly intractable in SPSS, whereas while it was a non-trivial exercise in SAS, it was certainly conceptually straight forward. And the resulting files could be analyzed in either R or SPSS, of course.(or SAS – which isn't a package that they licence because of its increasingly pricey )

Reply



**Bob Muenchen** *says:*

June 3, 2012 at 8:10 am

I've done quite a lot of complex data management in SAS, SPSS and R. To me they seem quite similar in capability except that R must fit the data into the computer's main memory (unless you're using Revolution Analytic's version). Where SAS may have the edge is reading unusual files where you have to read some data and, based upon that data, decide what other data to continue reading. I see that type of data rarely and I've only read it in

SAS. The others may be able to do it but I haven't taken the time to see if they can or not.

Reply

**Christian** *says:*

October 5, 2012 at 6:48 am

"To me they seem quite similar in capability except that R must fit the data into the computer's main memory"

I've been thinking about this lately, and I wonder if this might be a blessing in disguise? Every time our group hits memory constraints, we buy more RAM. It's cheap, and it grows exponentially cheaper/larger over time. Of course, that doesn't work for "very large problems". But, on the other hand, there's the MapReduce paradigm of divide-and-conquer. I don't often encounter datasets that I can't subdivide and process in chunks. Working with on-disc data is orders of magnitude slower (though SSD seems to help quite a bit), and so the dataset-in-RAM paradigm strikes me, after some thought, as a "good idea in disguise".

**Christopher D. Long (@octonion)** *says:*

March 25, 2013 at 6:36 pm

See the bigmemory package for R:

http://cran.r-project.org/web/packages/bigmemory/index.html

**Jeremie** *says:*

June 5, 2012 at 6:33 am

Excellent summary, thank you very much. The exponential growth of R packages is impressive.

I am trying to catch how you measured the statistical softwares on the job market.

Indeed a research with just "R" leads of course to nothing meaningful. I would search for expressions like theses :
"STATA (statistic OR statistical)" = 627
"MINITAB (statistic OR statistical)" =1277
"SPSS (statistic OR statistical)" = 2488
"R (statistic OR statistical)" = 2957
"SAS (statistic OR statistical)" = 7053

which shows the prevalence of SAS, but to a less degree.

Reply

**Bob Muenchen** *says:*

June 5, 2012 at 8:18 am

Many of the strings are easy:

JMP, BDMP Minitab, SPSS, Stata, Statistica, Systat

And SAS isn't too bad but but you have to exclude any hard drive interface references for which SAS has another meaning:

SAS (excluding SATA, storage, firmware)

R is devilishly difficult to get. Since you found more jobs for R than for SPSS I'm pretty sure you're getting mostly bad hits. You have to study a lot of the job descriptions to see what's actually being found. Plain old "R" is found in many irrelevant situations. I use a Linux shell script that searches for:

("SAS or R" or "R or SAS") and it repeats that pattern for the above packages and MATLAB, SQL, Java, Python, Perl

After much study that is the only way I have found to locate "R" that is relevant. If you find another way, I'd love to hear it!

The whole thing is a Linux shell script written by a former research assistant. A variation of it which I used for figures 7a and 7b is described in detail at:

http://librestats.com/2012/04/12/statistical-software-popularity-on-google-scholar/

Reply

**Jakob** *says:*

February 18, 2014 at 7:42 am

Another option to exclude lists is to manually inspect N samples of each query and estimate the chance of a query to be relevant. For example, you may get 5000 hits on an R query and estimate 1/20 to be referring to the statistical software -> approx 250 hits.

**Bob Muenchen** *says:*

February 24, 2014 at 3:17 pm

Hi Jakob,

I ended up doing something similar as described in How to Search for Analytics Jobs. I'll update the post to reflect this new perspective 2/25/14.

Cheers,
Bob

**omar** *says:*

June 17, 2012 at 3:13 pm

thank you for this stats article just what i needed

Reply

**rjrich** *says:*

July 8, 2012 at 5:11 pm

It would be interesting to include popular scientific plotting and statistics packages such as Origin Pro, SigmaPlot, and GraphPad Prism.

Reply

**Bob Muenchen** *says:*

November 19, 2012 at 10:21 am

Nice idea! However, I keep pretty busy collecting the current data.

Reply

**Ken** *says:*

July 10, 2012 at 11:48 am

Where you say, “No other data analysis languages covered by this article even make their top 100.”, is not true. If you look at the portion that says the next 50, covering 51-100 you will see S, S-PLUS, and SPSS which are all data analysis languages. It is also debatable that MATLAB, PL/SQL and Transact-SQL could be considered data analysis languages.

Reply

**Bob Muenchen** *says:*

July 10, 2012 at 4:10 pm

Ken, thanks very much for pointing that out. One of the hardest things about tracking so many sources of information is noticing all the changes that are relevant! I’ll deleted that sentence.

Reply

**Karup Pekar** *says:*

July 13, 2012 at 8:42 am

This is a very good article. I especially admire the way you have tried to quantify various measures. It’s worth reading just to learn that you can use “not” operators on google and amazon. Most illustrative of trends in stats packages and languages. Thank you!

Reply

**1273 ETT Result** *says:*

July 15, 2012 at 1:05 pm

Thankfulness to my father who shared with me regarding this website,
this webpage is genuinely remarkable.

Reply

**Ken** *says:*

September 25, 2012 at 1:15 pm

One interesting thing to look at could be comparing trends from the kdnuggets polls. You have the current year but there is also links to some of the prior years. For instance the following show two very different perspectives from two different points in time.

http://www.kdnuggets.com/polls/2011/tools-analytics-data-mining.html

http://www.kdnuggets.com/polls/2008/data-mining-software-tools-used.htm

I am not sure what all could be done with this but it would be interesting.

Reply

**Bob Muenchen** *says:*

October 10, 2012 at 7:41 pm

That's good idea. I'll do it if I can find the time!

Reply

**Monica Lewis** *says:*

October 6, 2012 at 5:12 pm

I'm curious about a review of tools used by non-statisticians for analysis in business. Do you know what products that help smooth some of most basic data related tasks that the masses are currently doing in Excel — such as pivot tables, commenting and collaboration? I've been building one to try to answer this, and am curious about others!

Thanks for all of the details on tool functionalities and preferences for true big data analysts!

Reply

**inundata** *says:*

November 16, 2012 at 3:44 pm

Fantastic! Is this published somewhere peer-reviewed that I can cite? I'm working on a journal article (which strongly discourages citing webpages) and would love to cite this as a source.

Reply

**Bob Muenchen** *says:*

November 19, 2012 at 10:18 am

Sorry, it's only available on this web site. I've had editors ask me to submit it, but I prefer to keep it as a living document that changes with the data.

Reply

**seo ranking tools** *says:*

November 29, 2012 at 10:43 pm

Hello! Do you use Twitter? I'd like to follow you if that would be okay. I'm absolutely enjoying
your blog and look forward to new posts.

Reply

**Bob Muenchen** *says:*

November 30, 2012 at 8:12 am

I'm @BobMuenchen on Twitter and I do tweet when each new post or article is finished. It's certainly OK to follow me. I don't tweet a lot, so you won't be bombarded with crazy messages about where I'm eating lunch!

Reply

**jergreen@gmail.com** *says:*

February 5, 2013 at 9:34 pm

SAS just doesn't seem affordable except for corporations. Do they even have a single user academic perpetual license?

Reply

**Bob Muenchen** *says:*

February 6, 2013 at 2:40 pm

SAS Institute never does perpetual licenses. A single user academic license is very expensive but they do make it very cheap per copy when you get an unlimited-copies license.

Reply

**Bob McConnaughey** *says:*

February 5, 2013 at 11:48 pm

I have quite a wonderful "ANCIENT" book that has a comparison of Stats/database packages circa 1980 back in my office. I DO remember that back in the day….the yearly license for the "Statistical Analysis Software" package was $1000.00 for a university. If I could attach a pdf I actually scanned the chapter on "General Statistical Packages." The book was basically the result of a survey of users…My favorite line: "More importantly, SAS's users think almost as highly of this program as its developer does"

Reply

**Bob Muenchen** *says:*

February 6, 2013 at 2:42 pm

I use SAS & SPSS frequently for clients and like them both. But when I get to choose, I usually use R.

Reply

**Dr.Az** *says:*

February 17, 2013 at 8:34 am

lovely post.
one tiny error– there are two captions titled the same serial "7a".
maybe you mean 7b in the latter one.

Reply

**Bob Muenchen** *says:*

February 21, 2013 at 3:24 pm

That's fixed now. Thanks very much for reporting it!

Reply

**Sue Briggs** *says:*

March 16, 2013 at 11:50 pm

"quiet" under Fig. 1d should be "quite"

Reply

**Bob Muenchen** *says:*

March 18, 2013 at 8:15 am

It's fixed. Thanks!

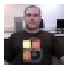
Reply

**ipad repair kl** *says:*

March 22, 2013 at 10:33 am

Hello! I know this is kinda off topic however I'd figured I'd ask.
Would you be interested in exchanging links or maybe guest writing a
blog post or vice-versa? My blog covers a lot of the same topics as yours and I believe
we could greatly benefit from each other.
If you might be interested feel free to shoot me an e-mail.
I look forward to hearing from you! Superb blog by the way!

Reply



**gawbul** *says:*

March 25, 2013 at 5:42 am

I'd love to see how Julia (julialang.org) fairs over the coming years

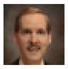
Reply

**Rosaria** *says:*

May 20, 2013 at 12:46 pm

Do you think you can include more of KNIME in some of your graphs? I am curious to see how it compares. I use KNIME and I have seen it cited only in figure 3 and figure 4.

Reply



**Bob Muenchen** *says:*

May 20, 2013 at 12:55 pm

Hi Rosaria,

I started out studying just classic statistics packages while the data mining software came from data collected by others. However I do hope to expand the graphs next year to include them. There's little real difference between the two types of software other than the user interface, which is better on most data mining packages.

Cheers,
Bob

Reply

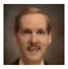
**Fred** *says:*

May 31, 2013 at 8:37 am

This is absolutely amazing. Given the passion that most scientists have towards their software packages and that you are a self-proclaimed Stata user, I'm amazed that you can have such an unbiased and rational approach to answering this question.

1) There seem to be way too many stats packages.
2) I was happy to see Number Cruncher Statistical Analysis in there. The copy I have is 10 years old, but I still use it for 3d graphing capabilities.
3) I conducted a web search of "SAS vs Stata" because a coworker uses Stata and won't shut up about it. I use SAS/Excel…and won't shut up about it. My hypothesis was that my coworker is using an outdated stats package and he is stubbornly set in his outdated ways. This article mostly disproves that hypothesis, but does give me some ammo on the comparison. Thanks!

Reply



**Bob Muenchen** *says:*

May 31, 2013 at 1:23 pm

Hi Fred,

I actually use Stata only occasionally, and then usually just to study how it does a particular thing. My co-author Joe Hilbe is the Stata guru. It is a beautiful system though. You can tell that a tiny number of people cared about making its structure consistent. SAS, SPSS and especially R were at the mercy of too many developers so their syntax is less consistent. All four are wonderful packages though, and each has an audience that thinks it's the best by far. I like 'em all!

Cheers,
Bob

Reply

**Wayne** *says:*

July 17, 2013 at 9:21 am

Bob,

I've used R for years, and just bought Stata/IC 13 yesterday for several reasons. First, the company has a great attitude/culture and it's always good to deal with a company where you like the people. Second, it seems to me to be the best option among SAS, SPSS, Minitab, et al, and it's also a better deal for an individual purchaser. Third, it implements some algorithms that are more advanced than the R equivalents. And fourth, Stata 13 was just released and has a lot of nice new features.

My first thought is that it reminds me a lot of Igor Pro, by Wavemetrics, which I used to use. They both have a great bunch of people (developers and users), a great culture, an interface that you can drive via commands or a GUI (though the GUI generates the command line equivalents so you can learn it or reuse it), and a consistent flavor. The difference being that Stata is statistics-oriented, while Igor Pro is scientific/experimental-oriented.

I like Stata a lot, but it won't replace R. I'd say that it's much more elegant than SAS, et al. (SAS was developed for punched cards and influenced by IBM's punched-card JCL, and has all kinds of obvious seams between its various parts. It's definitely a Frankenstein.) I'd disagree with you that Stata' syntax is more consistent than R's though. I believe you're talking about how functions in R have been written by various people, so the function calls may have some inconsistent argument names or perhaps result formats. On the other hand, Stata suffers from the data (essentially a spreadsheet) versus free-form variable (r(), e(), _b, _se, etc) distinction, which itself sets up various inconsistencies and makes me feel claustrophobic.

So I still think that R's the best option, but have definitely added Stata to my toolbox and it will be there long-term. I'd definitely recommend it to others.

To some degree, I think it makes a difference what direction you come to statistics from. If you're used to programming and like having the full machinery and flexibility of a programming language, R makes a lot of sense. If you don't really program — you just want to give commands and get results — though you want the option of automating some things or using programs that others have written, Stata makes a lot of sense.

**Bob Muenchen** *says:*

July 22, 2013 at 9:44 am

Hi Wayne,

Thanks for your interesting comments. I've talked to a couple of other people recently make the point that R is better as a programming language, while Stata is easier to use as a way to control pre-written procedures. SAS certainly has some odd inconsistencies, but I've used it for so many years that they seem second nature to me.

Cheers,
Bob

**Fred** *says:*

October 31, 2013 at 11:57 am

Hi Bob,

As I mentioned before, my coworker uses Stata and I use SAS/Excel. Unfortunately, my coworker has retired and I have no way to validate my SAS/Excel code with her Stata output. If I were to provide the code that my coworker used, datasets, and any other information required, could you reference somebody to me who can run a Stata program? I can't seem to find anybody who runs Stata!

Any information is helpful.
Thanks!
"Fred"

**Karl Rexer** *says:*

June 7, 2013 at 9:28 am

Great analysis, as always. This is a great resource for the entire analytics community. Thanks!

Reply

**Bob Muenchen** *says:*

June 7, 2013 at 11:01 am

Hi Karl,

Thanks very much! I really look forward to seeing your survey results each time. Keep that data coming!

Cheers,
Bob

Reply

**Kamal** *says:*

August 25, 2013 at 11:39 am

Hi Bob,
Thanks for providing an overall big picture of statistical packages. I am using SAS from past couple of years and is preparing for its certification too. As a beginner I always used to wonder about the differences among different statistical packages but your article has answered a lot of my questions.
Thanks.

Reply

**Bob Muenchen** *says:*
August 30, 2013 at 8:55 am

Hi Kamal,

I'm glad you found it useful.

Cheers,
Bob

Reply

**Ajay Ohri** *says:*
October 8, 2013 at 10:45 am

so why does SAS Institute still make 2.5 $ billion every year. Your data is overwhelmingly conclusive- but the SAS revenue is what makes me a hold out believer

Reply

**Bob Muenchen** *says:*
October 10, 2013 at 8:49 am

Hi Ajay,

As far as I know, SAS Institute is still the largest privately held software company in the world and I don't see that changing anytime soon. They continue to innovate, especially by offering complete solutions to problems rather than just offering tools that let you come up with your own solutions. I think the whole analytics pie is getting much larger. While SAS gets a smaller slice of this pie each year, it still adds up to more revenue.

Cheers,
Bob

Reply

**jamesmclarkJim Clark** *says:*
January 17, 2014 at 12:17 am

Nice data on use of different packages. A couple of comments. It would be interesting to know who is using what software and for what purposes. As an experimental psychologist, for example, I very much like SPSS for its handling of analysis of variance (both GLM and the older Manova). When I was generating course evaluations on my campus for a number of years, I liked to use SAS because of its powerful relational database functions (SQL). The same things could be done in SPSS but not nearly as "elegantly." Is it perhaps the case that different classes of users are finding the features they need in particularly packages? Finally, we might like to believe that the "best" product wins out, but that is not always the case with respect to software (e.g., Word vs Wordperfect?) and should perhaps warrant some caution with respect to usage statistics. Nice job!

Reply

**Bob Muenchen** *says:*

January 18, 2014 at 2:22 pm

Hi Jim,

You make some good points. Different packages definitely dominate in different market segments. Our campus (University of Tennessee) has a large social science presence and SPSS dominates by far overall. However, among economists Stata is dominant, the agriculturalists and business analytics folks use SAS, and while R use is in the minority, it seems like every department has someone on the cutting edge of their field using R.

I like all these packages for their various strengths and agree that it makes little sense to say which is "best" for everyone.

Cheers,
Bob

Reply

**Hypersphere** *says:*

February 26, 2014 at 3:32 pm

Extremely engaging. Although Sage is much more than a statistical package, it encompasses statistics, and it would be interesting to include it in the mix.

Reply

**john painter** *says:*

March 7, 2014 at 7:37 am

Should the caption for figure 1a say, "MORE popular"? The caption appears same as the one for Figure 1b

"Figure 1a. The number of analytics jobs for the less popular software"
"Figure 1b. The number of analytics jobs for the less popular software"

Your descriptions of the challenges faced when compiling data from readily available but harder to interpret data shows how much work you have put into this site. Thanks!

Reply

**Bob Muenchen** *says:*

March 7, 2014 at 8:08 am

Hi John,

Thanks for catching that! It's fixed. Regarding the amount of work, I wish I had tracked it. I do know that the job search section alone took over 100 hours. Now that I understand the problem better, I can update the figures in about an hour, but determining the optimal searches was really difficult.

Cheers,
Bob

Reply

**Joseph Hilbe** *says:*

March 17, 2014 at 9:01 pm

The primary reason I show either both Stata and R code or just R code for the examples in my books now is due to the fact that the far majority of statistics journal manuscripts that I referee or edit use R for examples. SAS and Stata seem to come in as the second most used stat packages. However, I realize that this may in part be due to the type of manuscripts I referee. I'm on the editorial board of six journals, and am asked to referee by a number of others. But these are generally related to biostatistics, econometrics, ecology, and recently astrostatistics (where Python and R are most common). It also seemed to me that most of the books I read or referenced when researching for my books also used R for examples, followed by Stata and SAS.

The second reason is due to the students I teach with Statistics.com. I teach 5 courses (9 classes a year) with the company. These are month long courses over the web with discussion pages which I use to interact with those enrolled in the courses. A good 95% (seriously) of enrollees are active researchers working in government, research institutions, hospitals, large corporations, and so forth, as well as university professors wanting to update their knowledge of the area, or learn about it if they knew little before. Students come from literally everywhere — the US, UK, Italy, Australia/NZ, Brazil, China, Japan, South Africa, Near Eastern nations, Nigeria, and even Mongolia. I always ask for their software preference, and have on average 15-30 students. Logistic Regression is the most popular course followed by Modeling Count Data. R is by far the most used software package. I started teaching with Statistics.com their first year (2003) , using Stata. I would accept submissions using SAS and SPSS, but the course text and handouts I used were in Stata. It is a very easy package to learn and it has a very large range of statistical capabilities. But I increasingly had more and more students wanting to use R. So I started to become more proficient, co-authored R for Stata Users with Bob Muenchen, (2010) which really spiked my knowledge of the software and now used the two package equally. My "Methods of Statistical Model Estimation" book with Andrew Robinson (2013) is a book for R programmers, and "A Beginners Guide to GLM and GLMM using R" (2013) with Alain Zuur and Elena Ieno uses only R and JAGS – I am ever more becoming a Bayesian as well. Other book, like my "Modeling Count Data" (Cambridge Univ Press) which comes out in May uses both R and Stata in the text, with SAS code for the examples in the Appendix. R, JAGS, and SAS is used for the Bayesian chapter.

Look through the new books that are being authored and the journal articles being published by the major statistics journals. Its mostly R, Stata, and SAS, with SPSS also used in books/journal articles specifically devoted to the social sciences. Minitab occasionally as well. Python and R almost exclusively for the physical sciences. For the many new books on Bayesian modeling, most use WinBUGS/OpenBUGS and R (and R with JAGS), and some SAS. I see Python becoming more popular though.

For what its worth, I've seen a lot of software over the years, From 1997 to 2009 I was Software Reviews Editor for The American Statistician, and received free stat software to review and use for 12 years and pretty much still ongoing. I turn 70 this year, so have watched the development of statistics and statistical software for quite awhile. I would not purchase stock in SPSS, nor in SAS for that matter. SAS is ingrained in the pharmaceutical and healthcare industry, and in much of "big" business, folks have jobs as SAS programmers, or SAS analysts. Too much is invested by business to simply drop it. But that's not the case as much with SPSS. With more Revolution-like businesses developing in the next decade, I believe R will predominate as the Franca Lingua statistical software. Stata will become ever more popular, but needs to develop a strong Bayesian component. Its not difficult to do given Stata's excellent programming and matrix languages. Python, OpenBUGS (WinBUGS is not being developed any more), JAGSs and perhaps some other Bayesian software will grow fast in use as well. The Predictive Analytics movement is having an influence as well, and together with academia is focused on employing more Bayesian, basic sampling, and enlightened machine learning into the analysis community.

Reply

**Bob Muenchen** *says:*

March 19, 2014 at 9:22 am

Hi Joe,

It’s good to hear from you! I, too, have noticed the rapid growth of R used as code examples in journals and books. I only measure books that use the software name in their titles since they’re easy to find. However, I do think it would be much more indicative of R’s dominance to somehow count the books that used R in examples. I see some that use R and Stata, or R and SAS, etc. so R may already be the dominant software used across all stat books.

Cheers,
Bob

Reply

**Vijayan Thankappan** *says:*
August 14, 2014 at 1:37 pm

Dear Sir,
The page linked below describes the capabilities of four different statistical software, and was intrigued to see a rather different take on capabilities of Stata. I love Stata, and it is a great tool to do routine in-built type analysis, but may its programming abilities are not that great?

http://stanfordphd.com/Statistical_Software.html

**Partha** *says:*
April 26, 2014 at 1:46 am

I am getting addicted by your writings. I am a student of statistics and want to learn as much as possible from your writings.

Reply

**Bob Muenchen** *says:*
April 26, 2014 at 7:44 am

Hi Partha,

I’m glad you’re enjoying them. It motivates me to keep working!

Cheers,
Bob

Reply

**David** *says:*
May 18, 2014 at 10:37 am

Excellent article, very detailed presentation of data. Good to follow the analytics trend. Thank you for this article.

Reply

**Simon** *says:*
May 20, 2014 at 9:54 am

Nice article, and impressive thinking too!

Just one query: In fig. 1a (2/2014), don’t you mean over 250 jobs, not under?

Reply

**Bob Muenchen** *says:*

May 20, 2014 at 1:39 pm

Hi Simon,

Thanks for catching that typo! It's fixed.

Cheers,
Bob

Reply

**Isaac** *says:*

June 3, 2014 at 2:54 pm

hello all,

I am a graduate of statistics. i want to focus my career in customer insight analysis, building predictive models. i have. learnt sql and SAS programming for data extraction and manipulation. I am confused on which stat package to really learn for data mining. I know SAS EM but feel coys wont employ based on point and click. What about SPSS? I would love to learn SAS programming for data minin on BASE SAS. Pls any recommendations as well as. books to read? thanks a lot.

**Jonathan Gezos** *says:*

September 2, 2014 at 6:01 pm

This is great, but you're missing out on a lot by only looking at tools that are 10+ years old. There is a lot of innovation in the industry right now with new players like Tableau and Looker in the mix.

Reply

**Bob Muenchen** *says:*

September 3, 2014 at 1:35 pm

Hi Jonathan,

Tableau is shown in figures 2a, 6b, 6d and 6e. However, most of those are from people who collected their own data. I'm focusing on advanced analytics or predictive analytics. Tableau is more of a visualization package. It does a nice job with a small number of variables but you can only see perhaps 8 at a time on a graph (x, y, z, color, size, shape, small multiples, time in animation). Even with that many it's hard to absorb. The other software can find patterns in hundreds or even thousands of variables. I just edited the paper to make the focus more clear.

Cheers,
Bob

Reply

**r4stats.com**

*The Twenty Ten Theme.* *Blog at WordPress.com.*