

Valparaiso University ValpoScholar

Business Faculty Publications

College of Business

5-2015

Selection of Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities

Ceyhun Ozgur
Valparaiso University

Michelle Kleckner
Valparaiso University

Yang Li
Valparaiso University

Follow this and additional works at: http://scholar.valpo.edu/cba_fac_pub


 Part of the [Business Commons](#)

Recommended Citation

Ozgun, Ceyhun; Kleckner, Michelle; and Li, Yang, "Selection of Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities" (2015). *Business Faculty Publications*. 22.
http://scholar.valpo.edu/cba_fac_pub/22

This Article is brought to you for free and open access by the College of Business at ValpoScholar. It has been accepted for inclusion in Business Faculty Publications by an authorized administrator of ValpoScholar. For more information, please contact a ValpoScholar staff member at scholar@valpo.edu.

Selection of Statistical Software for Solving Big Data Problems: A Guide for Businesses, Students, and Universities

SAGE Open
 April-June 2015: 1–12
 © The Author(s) 2015
 DOI: 10.1177/2158244015584379
 sgo.sagepub.com


Ceyhun Ozgur¹, Michelle Kleckner², and Yang Li¹

Abstract

The need for analysts with expertise in big data software is becoming more apparent in today's society. Unfortunately, the demand for these analysts far exceeds the number available. A potential way to combat this shortage is to identify the software taught in colleges or universities. This article will examine four data analysis software—Excel add-ins, SPSS, SAS, and R—and we will outline the cost, training, and statistical methods/tests/uses for each of these software. It will further explain implications for universities and future students.

Keywords

big data, Excel, Minitab, R, SAS, SPSS, statistical software

Introduction

Welcome to the age of big data, a revolutionary era where technology has transformed how businesses make decisions. According to a report by the McKinsey Global Institute, a trusted advisor for many influential businesses, “decision making will never be the same; some organizations are already making better decisions by analyzing entire data sets from customers, employees, or even sensors embedded in products” (Manyika et al., 2011, p. 5). In addition to intuition and judgment, businesses now use various software to draw conclusions from data sets and to thereby make decisions.

Surprisingly, schools do not teach students the same software that businesses look for. In his article that measures the popularity of many data analysis software, Robert Muenchen notes that discovering the software skills that employers are seeking would “require a time consuming content analysis of job descriptions” (Muenchen, 2014). However, he finds other ways to figure out the statistical software skills that employers seek. One of these methods is to examine which software they currently use. Muenchen includes a survey conducted by Rexer Analytics, a data mining consulting firm, about the relative popularity of various data analysis software in 2010. The results of the survey are pictured in Figure 1. As seen, data miners use R, SAS, and SPSS the most. Because 47%, 32%, and 32% of respondents use R, SAS, and SPSS, respectively, it can be inferred that these are the software skills that the greatest proportion of employers will continue to look for. However, this method only examines the software that employers might seek if they are hiring, so it does not accurately measure the software that they

currently look for. Muenchen's other method does this, studying software skills that employers currently seek as they try to fill open positions. In this approach, Muenchen puts together a rough sketch of statistical software capabilities sought by employers by perusing the job advertising site, Indeed.com, a search site that comprises the major job boards—Monster, Careerbuilder, Hotjobs, Craigslist—as well as many newspapers, associations, and company websites (Muenchen, 2014). He summarized his discovery in Figure 2.

As seen—in contrast to R's greater usage by companies over SAS, illustrated in Figure 1—job openings in SAS substantially lead open positions that require any other data analysis software. For employers, SPSS and R skills finish in second and third place. This second estimation method of Muenchen measures the software skill deficits in the job market. It seems that the demand for people with SAS skills outweighs the number of individuals with this capability. One reason for this disconnect could be that colleges and universities are not teaching SAS skills in proportion to the demand for these skills.

The convenience sample is collected based on school's answering our questionnaire. There was no other way to collect the data. We had to rely on the questionnaire that we

¹Valparaiso University, Valparaiso, IN, USA

²University of Michigan, Ann Arbor, USA

Corresponding Author:

Ceyhun Ozgur, Valparaiso University, 1909 Chapel Drive, Valparaiso, IN 46383, USA.

Email: ceyhun.ozgur@valpo.edu



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (<http://www.creativecommons.org/licenses/by/3.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (<http://www.uk.sagepub.com/aboutus/openaccess.htm>).

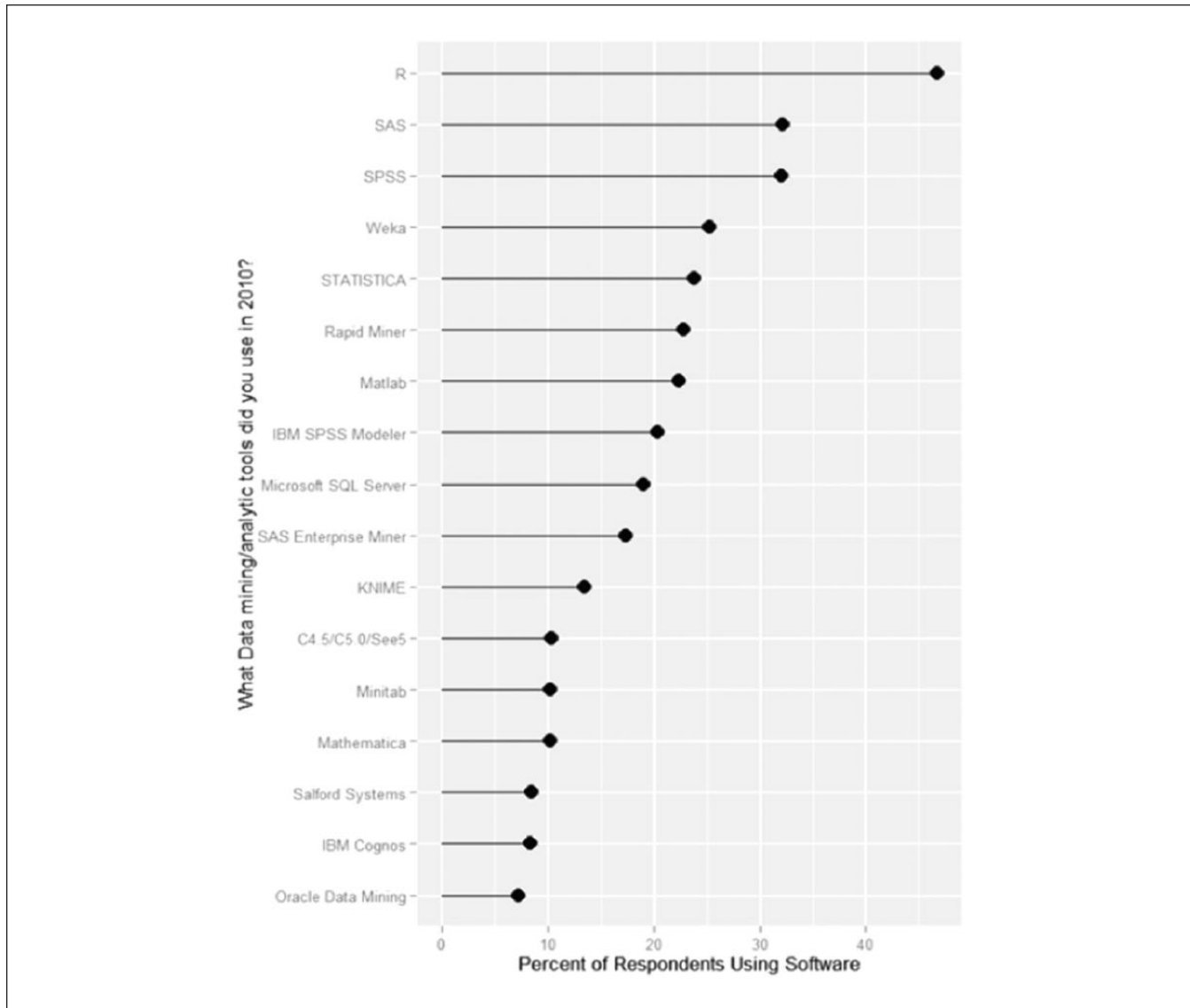


Figure 1. 2010 Rever analytics survey results of analytic tools.
Source: Muenchen (2014).

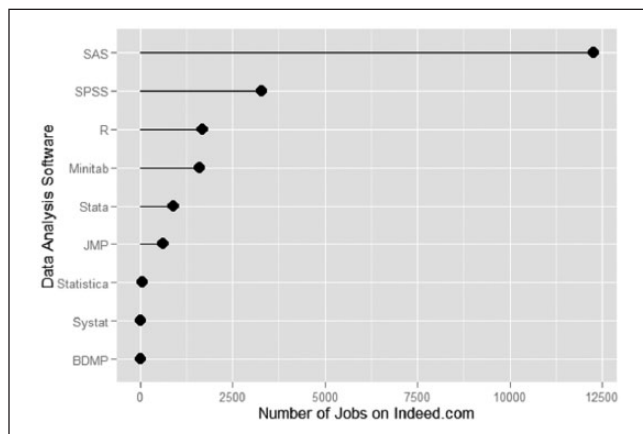


Figure 2. Jobs requiring various software.
Source: Muenchen (2014).

sent to establish this connect between software used and software sought. To assess this potential disconnect, we surveyed 18 departments—small and large, state and private, undergraduate and graduate, East and West—and, as expected, we discovered a discrepancy between the software taught at universities does not reflect this dominance. Only a few more departments teach SAS than R or SPSS. Some departments do not teach any software at all! Although this survey has a small sample size, the results do roughly sketch a trend seen in quantitative, engineering, and business departments across the country. College and university departments have not aligned their use of statistical packages to the skills demanded by employers. In many departments, SAS and

Table 1. Results From a Survey of Statistical Software Packages Taught.

School	Department	Software taught at graduate level	Software taught at undergrad level
Large, Midwestern, State University	Actuarial science	SAS, Excel, Mathematica	SAS
Medium, Southeastern, Private University	Biostatistics	SAS, SPSS, Minitab, Mathematica, Fortran, StatExact, Spatial Stat, C, C++	No Undergraduate Program in Biostatistics
Small, Southern, Private University	Computational and applied mathematics	Matlab, C, C++	Matlab, C, C++
Small, Midwestern, Private College	Mathematics	No graduate school	SPSS, Excel, Minitab, Mathematica
Large, Midwestern, State University	Mathematics	None	none
Medium, Northeastern, Private University	Mathematics	SAS, R, JMP, Matlab, DataDesk, ActivStats ^a	SAS, R, JMP, Matlab, DataDesk, ActivStats ^a
Small, Midwestern Private University	Mathematics and computer science	N/A	SAS, Excel
Medium, Northeastern Private University	Statistics	SAS, R, Excel, Minitab, JMP, Matlab, Python	N/A
Small, Southern, Private University	Statistics	SAS, SPSS, R, Excel, JMP, Matlab, Mathematica, Stata	JMP, Stata
Large, Southeastern, State University	Statistics	SAS, R, SAS Enterprise Miner	SAS, R, JMP
Small, Midwestern, Private College	Statistics	No graduate school	R
Large, Southeastern, State University	Engineering	Excel, JMP, Matlab, Mathematica, Mathcad	SAS, Excel, JMP, Matlab, Mathematica, Maple, Mathcad
Large, Southeastern, State University	Economics	N/A	SAS, R, ForecastX, GRETL
Small, Midwestern, Private College	Economics	No graduate school	Minitab, GRETL
Large, Southwestern, State University	Economics	No graduate program in economics	SPSS, Excel, Stata
Large, Southwestern, State University	Information systems and decision sciences	SAS, SPSS, Excel, Megastat, JMP, SAP, Minitab, Matlab, Stata, Mathematica ^a	SAS, SPSS, Excel, Megastat, JMP, SAP, Minitab, Matlab, Stata, Mathematica ^a
Large, Midwestern, State University	Marketing	SAS, SPSS, JMP	N/A
Large, Midwestern, State University	Marketing	SPSS, Excel ^a	SPSS, Excel ^a

Source. Compiled by Kleckner (2014).

^aThese schools did not specify whether the software listed were for graduate or undergraduate students, so we assumed both.

SPSS training are offered as noncredit generating instructions and also as part of mentoring. The skills required to teach Excel and SAS are very different. The departments need to invest in teachers who are capable of teaching these specialized courses using Excel, SAS, SPSS, and R.

The use of Excel as a software is very different than use of SAS or SPSS, but the cost and training is significantly higher for SPSS or SAS. We can use Excel add-ins to deal with the deficiency in software. We cannot solve as big of problems with Excel as with SAS or SPSS. However, Excel can still be useful software, especially with add-ins for small businesses.

Paying attention only to job availability, it seems that many schools need to reconsider their software choice in favor of implementing SAS. Nevertheless, there are many factors to consider other than the popularity within the job market. Schools must also consider the cost and time effectiveness of incorporating each software into their curriculum. Furthermore, specific departments within the school should consider which software best fits their area of study. One aim of this article is to provide this necessary information by outlining the cost, training requirement, statistical techniques, and specific uses within industry of leading data analysis software.

This article will not only assist schools in this software choice, but it will also help businesses decide which software is best to bring them to the next level of capability as big data analysis has become less of a privilege and more of a necessity. As Manyika et al. (2011) state,

The impact of developing a superior capacity to take advantage of big data will confer enhanced competitive advantage over the long term and is therefore well worth the investment to create this capability. But the converse is also true. In a big data world, a competitor that fails to sufficiently develop its capabilities will be left behind. (p. 5)

In other words, big data can no longer be ignored because companies that take advantage of it are winning the race against their less-modern competitors. More companies must consider implementing this type of analysis to stay competitive, and in doing so, they will need to understand which software is most appropriate for their business. This article will help gather and condense the necessary information for this type of decision. For those progressive companies that already utilize software to realize their goals, they can reconsider their current software choice in light of this comprehensive information.

Medical research is often weakened by poor statistical practice, and inappropriate use of statistical computer software is part of this problem. The statistical knowledge that medical researchers require has traditionally been gained in both dedicated and ad hoc learning time, often separate from the research processes in which the statistical methods are applied. Computer software, however, can be written to flexibly support statistical practice (Buchan, 2000).

All in all, this article will focus on SAS, SPSS, and R software because both methods in Muenchen's study indicate that they are the three most competitively sought software in industry. The article includes a complete collection of the cost-effectiveness, training, uses, and specific uses within industry of each of these software. It will begin, though, with an investigation of Excel and its add-ins because of their cost-effectiveness, utility, and availability.

Excel Add-Ins

What Are Excel Add-Ins?

Add-ins are programs that add optional features and commands to the traditional capabilities of Microsoft Excel. Excel has created add-ins for a multitude of purposes: data analysis, presentation, investment, business, personal, utilities and productivity tools, and organization. Within data analysis, some of the most popular add-ins include the Analysis Toolpak, Solver, and MegaStat. While the Analysis Toolpak and Solver are free add-ins, MegaStat is not. This section will focus on MegaStat and its usefulness in industry.

Cost

MegaStat costs US\$13.50, and prospective users can purchase it on McGraw-Hill's website.

Training

With the focus on science, technology, engineering, and mathematics (STEM) fields in this day and age, chances are most people who have used Microsoft Excel at some point during their lifetime. Although they may not have used the statistical analysis tools of Excel, they have used it to swiftly perform calculations of data, or at least to organize data. The creator of MegaStat, Orris Burdeane, explains, "Since MegaStat looks and works like Excel, almost anyone could use it to generate some output with just a few minutes of training" (O. Burdeane, personal communication, January 29, 2014). After all, MegaStat has dialog and input boxes, buttons, and checkboxes that work largely the same as those in standard Excel. Therefore, the 53-page tutorial pdf—complete with a step-by-step process to using each test that MegaStat performs, and pictures at every step—will likely provide more than enough guidance for trainees to effectively use this software.

Statistical Method/Tests/Uses

As will be seen later in this article, MegaStat can execute many of the same jobs that more costly software do. According to the McGraw-Hill website, MegaStat can perform a multitude of statistical operations: descriptive statistics, frequency distributions, probability, confidence intervals and sample size, hypothesis tests, ANOVA, regression, time-series/forecasting, chi-square, nine nonparametric tests, quality control process charts, and generation of random numbers ("Megastat," 2014). The creator of MegaStat himself acknowledges that SPSS and SAS, specifically, have more advanced options, "especially in the area of multivariate statistics" (O. Burdeane, personal communication, January 29, 2014). However, he believes that "MegaStat can handle most things encountered by non-PhD statisticians" (O. Burdeane, personal communication, January 29, 2014).

The major caveat for this cheap and easy-to-use software is its size capability. For example, Burdeane experimented with the number of data points that MegaStat can handle a few years ago. "I did find a file with 10 columns and 152630 rows. That is over 1 ½ million data points and MegaStat did a descriptive statistics analysis on it in about 10 seconds" (O. Burdeane, personal communication, January 29, 2014). While the capability to analyze a million and a half data points sounds like a tremendous feat, and it is, this capability does not meet the demand of large companies like Wal-Mart and Facebook. An article published by SAS called "Big Data Meets Big Data Analytics" puts it plainly: "Wal-Mart handles more than a million customer transactions each hour and

imports those into databases estimated to contain more than 2.5 petabytes of data,” and “Facebook handles more than 250 million photo uploads and the interactions of 800 million active users with more than 900 million objects (pages, groups, etc.) each day” (SAS, 2014b). Extracting this data and making use of it using MegaStat is just not feasible.

Burdeane also mentioned a couple of other restrictions of MegaStat, including its limitation to 12 independent variables in multiple regression and further restrictions on variables and table size (O. Burdeane, personal communication, January 29, 2014). While MegaStat does have all of these constraints, the fact remains that it is an extremely powerful software given the cost and training required.

Specific Uses in Industry

The creator of MegaStat believes that this software is used by a different group of people from the major statistical packages. Burdeane (personal communication, January 29, 2014) states,

I would guess that most use of MegaStat in companies is by people who are not professional statisticians. I think people with formal training in statistics beyond an introductory course would have experience with one of the big packages (SAS, SPSS, Minitab) and would tend to stick with that software even if it was overkill for many analyses.

He suggests that many analyses do not require major packages, like SAS, SPSS, and R, but statisticians stick to them because they are comfortable.

However, companies within industry do still use Excel. For example, a global appliance manufacturer uses Excel “for extensive ‘What If’ analysis around budgeting” and to forecast (J. Ward, personal communication, January 20, 2014).

SPSS

What Is SPSS?

SPSS, originally termed Statistical Package for the Social Sciences, was released in 1968 as a software designed for the social sciences. Since then, IBM has replaced SPSS Inc. as the owner, and the software has expanded its user base past this one area. The software’s former acronym has been replaced with Statistical Product and Service Solutions to reflect the greater diversity of its clients. Arguably, it still remains the leading statistical analysis software package for the social sciences.

Cost

Obviously, consumers can buy SPSS software packages separately by choosing a particular product that they think will satisfy their need; however, SPSS offers bundles that cost

Table 2. Prices of Bundles Offered by SPSS.

Package	Features	Price
Standard	Authorized user license	US\$5,270
	Authorized user initial fixed term license	US\$2,320
	Concurrent user license	US\$13,200
Professional	Concurrent user initial fixed term license	US\$5,810
	Authorized user license	US\$10,600
	Authorized user initial fixed term license	US\$4,660
	Concurrent user license	US\$26,500
Premium	Concurrent user initial fixed term license	US\$11,600
	Authorized user license	US\$15,800
	Authorized user initial fixed term license	US\$6,950
	Concurrent user license	US\$39,400
	Concurrent user initial fixed term license	US\$17,400

Source: IBM (2014a).

much less than paying for the programs independently. SPSS offers three of these bundles: standard, professional, and premium.

Within each of these bundles, SPSS gives four options: an authorized user license, authorized user initial fixed term license, concurrent user license, and concurrent user initial fixed term license. Thus, when customers decide they want to purchase SPSS, they have to make two decisions: user license versus initial fixed term license and authorized user versus concurrent user. User licenses never expire, while initial fixed term licenses last for 12 months. An authorized user is a single licensee who buys the right to use the program; a concurrent user is the right for a single person to use the program at a given time, but it does not distinguish who this person has to be.

With these descriptions in mind, beneath in Table 2 are the prices for purchasing the three different bundles of SPSS.

SPSS also offers student packages for college attendees. Students can purchase the single user initial fixed-term license “SPSS GradPack” software from their college or university, or they can buy it from SPSS’s official distributors, like Creation Engine, On the Hub, StudentDiscounts.com, Studica, and ThinkEDU (IBM, 2014b). For example, on the Creation Engine website, students can buy the SPSS Statistics Premium GradPack for US\$98.95 (“IBM SPSS Statistics Premium GradPack 22,” 2014).

Training

In her article about the use of statistical software for sociology, Ashley Crossman addresses the difficulty—or lack thereof—of using SPSS for the first time. She explains,

SPSS provides a user interface that makes it very easy and intuitive for all levels of users. Menus and dialogue boxes make it possible to perform analyses without having to write command syntax, like in other programs. It is also simple and easy to enter and edit data directly into the programs. (Crossman, 2014)

On the surface, these descriptions make SPSS sound a lot like Excel. In fact, SPSS does look similar to typical spreadsheet applications like Excel, and its ease of use is very comparable to Excel as well.

Statistical Method/Tests/Uses

There are many differences between Excel and SPSS that suit SPSS to better handle statistical methods. For one, “SPSS was designed specifically for statistical processing of large amount of data at an enterprise level,” while spreadsheets are broadly applicable to many different tasks outside of statistical computing (Robbins, 2012). An advantage of this specialized design is that SPSS “keeps calculated statistics and graphs separate from the raw data but still easily accessible” (Robbins, 2012). SPSS software furthermore has a much more convenient platform for performing statistical tests. For instance, performing a one-sample *t* test in Excel requires some independent calculations by the user, whereas with SPSS, the user only needs to “select a variable and supply the value to compare with [the] sample” and click “Ok” (Robbins, 2012). Another advantage of SPSS is that it links numerically coded data to its original meaning (Robbins, 2012). With most data being electronically stored in numerical fashion, this feature of SPSS is highly valuable.

For these reasons, SPSS is well suited to statistical analysis, but what statistical procedures can SPSS handle? SPSS’s standard bundle includes its statistics base, advanced statistics, bootstrapping, custom tables, and regression capabilities. Purchasing the professional bundle further supplies the consumer with the categories, data preparation, decision trees, forecasting, and missing values features. The most comprehensive bundle, premium, provides the user with the complex samples, conjoint, direct marketing, exact tests, neural networks, amos, sample power, and visualization designer, in addition to all of the packages from the professional bundle (“SPSS Statistics,” 2014).

After the statistical analysis is complete, SPSS is also useful for generating plots of distributions and trends, charts, and tabulated reports.

Specific Uses in Industry

IBM’s SPSS software has spanned many industries. On its website, prospective clients can read about SPSS success stories in fields like automotive, banking, chemical and petroleum, computer services, consumer products, education, electronics, energy and utilities, and on and on. They can also access a list of SPSS’s clients, such as Barclays,

Kaplan, and Wimbledon Championships. Below are a couple more specific examples of SPSS at work within industry.

- Infinity insurance uses SPSS’s predictive analytics feature to detect fraudulent claims (IBM, 2014c).
- “By mining alumni and stakeholder records, social media and other unstructured data-sets with text analytics software, [Michigan State University] gains insights into the engagement, sentiments and behavior of current and potential donors,” which enables smarter fund-raising (“Success Stories for SPSS,” 2014).
- The Guardia Civil, Spain’s very first national law enforcement agency, has investigated crimes and psychology using SPSS (“Success Stories for SPSS,” 2014).
- One distinguished hospital uses SPSS to forecast payment behavior. It tries “to better identify patients who are most likely to pay their hospital bills” by what it calls “predict[ing] patient payment potential” (“Success Stories for SPSS,” 2014).

SAS

What Is SAS?

SAS (Statistical Analysis System) is a commercial statistical package that was developed during the 1960s and 1970s at North Carolina State University as part of an agricultural research project. Its usage has grown exponentially since then. Nowadays, 91 of the top 100 companies on the 2013 Fortune Global 500 list use the software (SAS, 2014a). This article will discuss the two main SAS starter packages: Analytics Pro and Visual Data Discovery.

An important fact to note about SAS is that the software does not run on Mac computers very easily (one way to run the software is through parallels, where users buy and run the Windows interface as well).

Cost

An individual license of the Analytics Pro version of SAS costs US\$8,700 for the first year and US\$2,436 for each year thereafter. The cost for each renewal is 28% of the amount originally paid in the first year. With a few more features than the Analytics Pro system, the Visual Data Discovery package costs US\$10,800 for the first year of use. Like Analytics Pro, renewing this package costs 28% of the original cost, so it costs US\$2,822.40 for each additional year. However, these prices only apply to customers working with their own data. If a user wishes to perform data analysis for the benefit of some other party, then he must secure a different license by consulting a SAS representative (SAS, 2014d).

One of these alternative licenses is a server-based license. These licenses certainly save schools and businesses money

by allowing their affiliates each to access the software through a web-based connection or a network. SAS fills these requests on a case-by-case basis, so interested customers should speak to SAS directly to get a quote (SAS, 2014d).

On top of these two versions, SAS has created an OnDemand edition, which is available at no cost to degree-granting institutions. Professors can set up an account online, and they and their students can access the software anywhere with an Internet access. Although this free software “has been reported to be slow at times,” it definitely provides a great opportunity for schools to teach students the basics of SAS programming (Loomis Lofland & Ottesen, 2013, 3).

While the proper versions of SAS do come at a steep financial cost, they furthermore cost time in the form of installation. Chelsea Loomis Lofland and Rebecca Ottesen speak to this expense in their article, “The SAS Versus R Debate in Industry and Academia.” They explain, “SAS can be difficult for users to obtain and the initial installation is sometimes tricky . . . long and difficult” (Loomis Lofland & Ottesen, 2013, p. 3). However, in contrast to some other software (like R, as will be seen later), SAS only requires this initial installation. It does not require users to install any packages in the future. Everything is included in this set-up.

Training

Ashley Crossman accurately advises, “SAS is a great program for the intermediate and advanced user because it is very powerful, can be used with extremely large data sets, and can perform complex and advanced analyses” (Crossman, 2014). After all, SAS requires more training than Excel and SPSS because it largely runs on programming syntax rather than point-click menus that other software boast.

The amount of training necessary for individuals to properly use SAS depends on many factors, including the trainee’s background and the type of analysis she will need to perform. In terms of background, prospective SAS programmers with prior programming experience will have a much easier time. SAS syntax resembles that of other programming languages, so experience with one language often helps learn another. For instance, SAS is similar to Java in that both contain data values, function calls, identifying key words at the beginning of each line, and semicolons at the end of each line (Boudreaux, 2003). But, even if the syntax of SAS and a previously learned language are completely different, experience with coding is extremely helpful because the art of programming is a different kind of thinking. It is easy to migrate from one tool to the other. The “primary” push is to get organizations to use proper tools. The training required also depends on the type of analysis that the trainee must carry out. If the trainee only needs to run the same type of test repeatedly, then she may only need training in a specific aspect of SAS programming; however, if the trainee will need to develop a process based on each new task, then she will need more sound understanding of the software.

Fortunately, experts have written copious texts about how to use SAS, and SAS has a strong user support system, so even if users do not have complete understanding of the software, they can run it. While there exists no easy way to calculate the number of books written about SAS, Robert Muenchen did as best he could to estimate these statistics by searching for books published with “SAS” in their title. He found that close to five hundred of these books were published between 2001 and 2011 (Loomis Lofland & Ottesen, 2013). In addition to all these useful texts, experts really cannot pinpoint any issues with the user support of SAS. Loomis Lofland and Ottesen clarify,

SAS has extensive online documentation, expert technical support, professional training courses, many excellent books in press, and a tight-knit user group and web-based community. Problems can be addressed to SAS directly via tech support who replies very quickly and will work with the user to solve the problem. (Muenchen, 2014)

They designate the user support service of SAS as one of its main specialties. Therefore, even though SAS requires some programming skills, the strength of SAS’s support system makes it more manageable for less advanced users.

Statistical Method/Tests/Uses

SAS’s Analytics Pro bundle comes with three of the most popular SAS products: Base SAS, SAS/STAT, and SAS/GRAPH. The corporation’s Visual Data Discovery collection includes SAS Enterprise Guide (SAS’s only point-click interface) and JMP software to make discovery and exploratory analysis easier.

With either of these toolsets, programmers can perform a number of statistical tests. The Institute for Digital Research and Education website outlines a multitude of statistical tests and their corresponding SAS codes. The list includes 32 tests that come from statistical categories such as regression, factor analysis, discriminant analysis, ANOVA, nonparametric tests, and correlation (UCLA: Statistical consulting group, 2014). The full list can be seen in Table 3.

SAS can perform many more statistical tests than just these, though. It also functions well with forecasting, time-series analysis, and many other advanced statistical techniques. In fact, SAS has created specialized programs for these methods. The SAS website’s “Products & Solutions” page has a complete list of these programs.

Also on this page, SAS has additional packages to access that are industry-specific. For example, there is a “SAS Drug Development” package that “enables the efficient development, execution, and management of analysis and reporting activities for clinical research”; a “SAS Fraud Management” package that “delivers a full-service enterprise-wide fraud management system that offers real-time scoring of accounts by looking at all card transactions—including purchases,

Table 3. List of Tests That SAS Can Perform.

One sample t test	One sample median test	Binomial test	Chi-square goodness of fit
Two independent samples t test	Wilcoxon–Mann–Whitney test	Chi-square test	Fisher's exact test
Kruskal–Wallis test	Paired t test	Wilcoxon signed rank sum test	McNemar test
One-way repeated measures ANOVA	Repeated measures logistic regression	Factorial ANOVA	Friedman test
Ordered logistic regression	Factorial logistic regression	Correlation	Simple linear regression
Nonparametric correlation	Simple logistic regression	Multiple regression	Analysis of covariance
Multiple logistic regression	Discriminant analysis	One-way MANOVA	Multivariate multiple regression
Canonical correlation	Factor analysis		

Source. UCLA: Statistical Consulting Group (2014).

payments and nonmonetary transactions”; and a “SAS Risk Management for Insurance” package that “implements the Solvency II standard model approach for calculating risk-based capital with [its] comprehensive solution for performing risk analysis and risk-based capital calculations” (SAS, 2014c). On top of these specialized packages for health care, banking, and insurance, SAS has formulated software with built-in functions for other areas like law enforcement, communications, retail, casinos, utilities, and sports, among others.

SAS's advantageous functions extend beyond just carrying out statistics, though. It has superior qualities for both before the statistical analysis and after. Prior to the actual statistics, it facilitates the reading in and managing disorganized data. Real-life data are rarely clean and analysis ready. SAS can interpret messy data sets, convert them to a clean form, and manipulate them in ways that other software cannot (Loomis Lofland & Ottesen, 2013). After the user performs the statistics, SAS has impressive graphics and report writing features that will help disseminate the findings in clear and appealing ways. But, these aesthetic products come with a caveat according to Loomis Lofland and Ottesen. They explain, “SAS provides many useful procedures for creating detailed and polished reports”; however, “some of the more detailed reporting procedures . . . have a learning curve that takes place before being able to use them correctly” (Loomis Lofland & Ottesen, 2013, pp. 3-4).

Specific Uses in Industry

As stated in the previous section, SAS has built-in, functional packages for many specific industries, including health care, banking, insurance, law enforcement, communications, retail, casinos, utilities, sports, and more. Below are a couple of real-life uses of SAS within some of these industries.

- A leading medical device company utilizes SAS “for clinical study data analysis” (K. Kleckner, personal communication, February 1, 2014). This same company furthermore uses the software “for setting sample sizes for pre-clinical studies and human clinical studies; [and] for setting controls on manufacturing operations” (K. Kleckner, personal communication, February 1, 2014).

- A global appliance manufacturer uses SAS for quality control by performing predictive analyses of product defects (J. Ward, personal communication, January 20, 2014).

R

What Is R?

R is a free, open-source statistical software. Colleagues at the University of Auckland in New Zealand, Robert Gentleman and Ross Ihaka, created the software in 1993 because they mutually saw a need for a better software environment for their classes. R has certainly outgrown its origins, now boasting more than 2 million users according to an R Community website (Revolution Analytics, 2014).

Cost

R is free and is downloadable from the Internet. To repeat, it has no subscription fees, user limits, or license managers. However, this presents a danger. As open-source software, R could be a security concern for large companies because the software can be freely used, changed, and shared by anyone.

Like SAS, R can be expensive in a form other than monetary. While the base for R is very easy to install, users must download packages to perform specific analyses, which can be very time-consuming (Loomis Lofland & Ottesen, 2013). For example, as of this article, there are 5,508 available packages, and this number grows weekly if not daily (Comprehensive R Archive Network, 2014). This provides many options, but searching through the assemblage of choices can be difficult and time-consuming. One of the strengths of R is its ability to find packages that have very specific solutions or we may generate solutions.

Training

Again, like SAS, the training necessary for effectively using R depends on the previous computing experience of the trainee. Computing experience is helpful because data analysis in R requires writing functions and scripts, not just

pointing and clicking like in Excel or SPSS. In many ways, though, R is comparable with other programming languages. For instance, similar to many other languages, it is a command line interface. In addition, its source code is similar to that of C and Fortran, and it supports matrix arithmetic and data structures like APL and MATLAB. Having used any of these in the past could lessen the training time necessary to learn R. As stated with SAS above, though, having any programming experience at all often will speed up the learning process for trainees as programming problems are a completely different type of puzzle.

Sources report varied answers when identifying the training necessary to successfully utilize R. Some believe that R does not necessitate much knowledge of computer programming after all. For example, Daryl Pregibon of Google testifies that R “allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems.” He is saying that packages abstract the programmer from the implementation of specific tools (decision trees/forests, etc.). Ashlee Vance of *The New York Times* informs, “R has quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use” (Vance, 2009). R is, after all, not as daunting as other languages, having very natural and expressive syntax for data analysis. The syntax is not very common to traditional programmers. Some programmers may even label its syntax as quirky. In R language, “`anova(object_1, object_2)`” produces an ANOVA table, “`coef(object)`” extracts the regression coefficient, and “`plot(object)`” produces plots showing residuals, fitted values, and other diagnostics (“An Introduction to R,” 2014). Still, R does require the use of objects, operators, and functions before applying these intuitive commands. Fortunately—as stated earlier—many packages are available for download and use off the Internet, so users do not necessarily have to know the code or write it. This is another reason why some people say that R does not require much programming knowledge.

However, because of errors in some of these packages and lack of user support for R, others believe that advanced training investment is necessary to use the software. Two people who hold this viewpoint are Loomis Lofland and Ottesen, who say,

[R] users rely on what others put out there about the software. . . . Packages are not written by the R Development Core-Team; therefore, they are not well polished and some could have questionable validity. It is also difficult to direct an issue to a particular person or support system. (Loomis Lofland & Ottesen, 2013, p. 3)

Although R may be usable without much coding experience, when a problem arises, the lack of programming knowledge will become evident and costly due to a dearth of documentation and technical support for resolving the issue.

In other words, people without sufficient knowledge of the R programming language can implement the syntax in their own use, but they do not necessarily have solid understanding of what the code actually says. This lack of R coding knowledge makes debugging difficult if not impossible, and it could lead to erroneous results with severe decision-making consequences.

Loomis Lofland and Ottesen also explain that report writing in R is difficult. They claim that the extensive programming required to code a report in R is quite a time investment, as “R does not have a defined way of producing reports” (Loomis Lofland & Ottesen, 2013, p. 3).

Statistical Method/Tests/Uses

R is a comprehensive statistical analysis toolkit. It can perform any statistical analysis desired, but users must either write the code or access the code from someone who has already written it. As stated on its website, people have already designed many standard data analysis tools “from accessing data in various formats, to data manipulation (transforms, merges, aggregations, etc.), to traditional and modern statistical models (regression, ANOVA, GLM, tree models, etc.; “Why Use R?” 2014)”. Actually, programmers have designed many more packages than just these. As stated earlier in this section, programmers have already coded 5,508 packages. These include packages for Bayesian statistics, time-series analysis, simulation-based analysis, spatial statistics, survival analysis, and many, many more (Comprehensive R Archive Network, 2014). For a complete list of packages already designed for R, visit <http://cran.us.r-project.org/web/packages/>.

The key feature of R that differentiates it from other statistical softwares is its acceptance of customization. On one hand, the aforementioned software have “data-in-data-out black-box procedures” (“Why Use R?” 2014). In other words, the developers have written the code for a certain function, such as performing decomposition for a time-series model, and users have never seen this built-in code that runs in the background. All they need to do is use a “decomp” command, or something of the sort, and the statistical package will perform the decomposition for them. However, R is an interactive language. It requires users to write the code (for the decomposition, or whatever procedure desired) or to paste the code in from someone who already wrote it. Because the function’s code is visible in their command box, users can manipulate the commands however they see fit. Thus, R enables experimentation and exploration by allowing users to improve the software’s code or to write variations for specific tasks. They can even mix-and-match models for better results. With the prepackaged functions in the other statistical software, this is not as easy.

After completing a statistical analysis with R, the software is known for generating appealing charts and tables. The custom charting capabilities of R create “stunning

infographics seen in *The New York Times*, *The Economist*, and the FlowingData blog” (Revolution Analytics, 2014).

With R, though, it is important to acknowledge that it cannot manage messy data as easily as other available statistical software. Loomis Lofland and Ottesen warn,

The design of R was focused around statistical computing and graphics, so data management tends to be time consuming and not as clean as SAS. . . . Students who have used solely R have an unrealistic expectation of the state of the data they receive. (Loomis Lofland & Ottesen, 2013, p. 3)

But, once the data is organized, as stated, R is an invaluable data analysis performer and graphics creator.

Specific Uses in Industry

The usage of R across diverse domains is undeniable. A *New York Times* article cites its practice in major companies, like Google, Pfizer, Merck, Bank of America, the InterContinental hotels Group, and Shell (Vance, 2009). For specific examples, see below.

- Google “taps R for help understanding trends in ad pricing and for illuminating patterns in the search data it collects” (Vance, 2009).
- Pfizer has engineered its own custom packages in R, which allows scientists to manipulate their own data during nonclinical drug studies instead of hiring a statistician to do the work for them (Vance, 2009).
- A financial services company utilizes dozens of R packages to perform derivatives analysis (Vance, 2009).

Conclusion

Excel add-ins are well suited to small companies and small projects because of their availability and low cost, while SPSS, SAS, and R work well for large projects and large businesses because of their ability to handle large sums of data efficiently. As discovered at the beginning of the article, Excel’s MegaStat option can execute many important statistical procedures that people trying to interpret smaller data sets can utilize for low financial cost and training cost. However, as stated, MegaStat can only manage a certain amount of data. Therefore, larger data sets, such as those accumulated by Wal-Mart and Facebook, require a higher powered software, like SPSS, SAS, or R. Differentiating between which of these software best fits the analysis of these larger data sets depends on a number of factors, and each statistical package has its own strengths and weaknesses. Hence, this article has investigated the cost, installation procedure, training necessary, built-in packages, level of desire to manipulate code, user support, importance of appealing graphics, and many other considerations in hopes

of providing businesses and universities with details that can ease their choices of software.

Implications

With the tremendous (and growing) focus on big data in today’s society, businesses, universities, and future students should actively participate. Big data are a mixture of structured and unstructured data and not just number of records. Big data capabilities go beyond the maximum number of rows and columns a software can handle. Big data consist of both structured and unstructured data; the McKinsey Global Corporation explains why. It states,

The use of big data will become a key basis of competition and growth for individual firms. . . . From the standpoint of competitiveness and the potential capture of value, all companies need to take big data seriously. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value from deep and up-to-real-time information. Indeed, we found early examples of such use of data in every sector we examined. (Manyika et al., 2011, p. 6)

Noncredit courses taught at Universities should also be included in the discussion of teaching these software, such as SPSS, SAS, and R. To repeat, businesses are leveraging big data in every sector that McKinsey examined. Hence, businesses need to take notice. McKinsey further explains,

There will be a shortage of talent necessary for organizations to take advantage of big data. [By 2018,] we project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions. (Manyika et al., 2011, p. 6)

In other words, McKinsey estimates that by 2018, the United States will lack at least 140,000 people with expertise in big data. Big data analysis does not occur by hand, so McKinsey essentially predicts that the United States could need this many more people with expertise in software that can handle big data. For this reason, colleges, universities, and future students should assume the preparatory measures necessary to combat this deficit. After all, future students with training from college and universities will need to fill these gaps.

This article aimed to help businesses, schools, and students recognize what they can do to improve their performance and utility in this data-driven society. To begin, from this article, businesses can learn the most effective tools to suit their context. New businesses can learn the tools that they should buy (or simply download), while older businesses can learn the software that could better serve their needs than the one they currently have. Finding the suitable software is important because companies that employ the most efficient data analysis software will compete better

against competition, by effectively accessing and using their stockpiles of data to make better decisions.

Colleges and universities could improve job placement by preparing students in the specific software that hiring companies use. As seen in the introduction, the majority of companies use the software reported in this article. Therefore, schools would benefit from teaching students to use these software.

Students can add a “software taught” category to their list of traits sought in higher education to prepare themselves for job placement. One of the most important decisions that future students make is selecting a major. Often, a student’s desired major can influence the selection set. However, other decisions are growing in importance too. In terms of finding a job, employers are increasingly seeking out recent graduates who have experience with big data software, like SPSS, SAS, and R. Therefore, it is becoming more important for students to seek out a university that will prepare them with knowledge of pertinent software, which will increase their likelihood of finding a satisfying job. Obviously, careers in big data will be abundant, so prepared students will have little trouble finding a job in that area. Nevertheless, students trained on high demand software will have more and better options for job placement.

Future Research

We plan to continue our research of software. While we acknowledge that no one software is suitable for every project within a specific sector of industry, we do believe that certain software may be best matched to the majority of projects within an industry. By further surveying the use of software by particular businesses, we hope to discover which software is best suited to business, mathematics, statistics, engineering, and other majors. This information should benefit businesses, schools, and students. We would also like to compare this with the software that academics use in their research.

Furthermore, we will study database software in much the same manner that we studied statistical software. As we surveyed businesses and universities about the statistical software that they utilize and teach, respectively, many of these establishments included database software in their responses. After all, database software serves as the means to organize the data sets, so the ability to work with database software is often just as important as the ability to analyze the data sets with statistical software.

Also, we will look at statistical software and their effectiveness in teaching. This article has focused primarily on big data software and their usefulness for finding a job within the business world. Another avenue to explore is the feasibility of using these software and others as tools for learning statistical concepts. Perhaps big data software such as SAS, SPSS, and R are extremely effective for business tasks, but they are not as effective for learning statistical concepts in the classroom. Statistical software that may better suit the classroom environment

include Minitab or Statistica. An analysis of the pros and cons for these software (and potentially others) is another future goal of ours.

Acknowledgments

We would like to thank the various businesses and universities that responded to our inquiries about software and therefore made this article possible.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

- An introduction to R. (2014). Retrieved from <http://cran.r-project.org/doc/manuals/R-intro.html#Statistical-models-in-R>
- Boudreaux, D. (2003, March-April). *Java Syntax for SAS programmers*. Paper presented at the SAS Users Group International, Seattle, WA.
- Buchan, I. E. (2000). *The development of a statistical computer software resource for medical research*. Liverpool, UK: University of Liverpool.
- Comprehensive R Archive Network. (2014). *Contributed packages*. Retrieved from <http://cran.us.r-project.org/web/packages/>
- Crossman, A. (2014). *Analyzing quantitative data: Statistical software programs for use with quantitative data*. Retrieved from <http://sociology.about.com/od/Research-Tools/a/Computer-programs-quantitative-data.htm>
- IBM. (2014a). SPSS Statistics. Retrieved from <http://www-01.ibm.com/software/analytics/spss/products/statistics/buy-now.html>
- IBM. (2014b). *SPSS Statistics GradPack*. Retrieved from <http://www-03.ibm.com/software/products/en/spss-stats-gradpack/>
- IBM. (2014c). *Why SPSS Software?* Retrieved from <http://www-01.ibm.com/software/analytics/spss/>
- IBM SPSS Statistics Premium GradPack 22. (2014). Retrieved from <http://www-03.ibm.com/software/products/en/spss-stats-gradpack>
- Kleckner, M. (2014). The Collection of Data for Table 1 Results From a Survey of Statistical Software Packages Taught.
- Loomis Lofland, C., & Ottesen, R. (2013). *The SAS Versus R debate in industry and academia*. Paper presented at the SAS Global Forum 2013, San Francisco, CA.
- Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Byers. “Big Data: The next Frontier for Innovation, Competition, and Productivity.” Big Data: The next Frontier for Innovation, Competition, and Productivity. McKinsey & Company, 1 May 2011. <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>.
- Megastat. (2014). Retrieved from http://highered.mheducation.com/sites/0070983755/student_view0/megastat.html
- Muenchen, R. A. (2014). The popularity of data analysis software. Retrieved from <http://r4stats.com/articles/popularity/>

- Revolution Analytics. (2014). *What is R?* Retrieved from <http://www.inside-r.org/what-is-r>
- Robbins, S. (2012). *How does SPSS differ from a typical spreadsheet application.* Retrieved from <https://publish.illinois.edu/commonsknowledge/2012/06/07/how-does-spss-differ-from-a-typical-spreadsheet-application/>
- SAS. (2014a). *About SAS.* Retrieved from http://www.sas.com/en_us/company-information.html
- SAS. (2014b). *Big data meets big data analytics: Three key technologies for extracting real-time business value from the big data that threatens to overwhelm traditional computing architectures.* Retrieved from http://www.sas.com/resources/white-paper/wp_46345.pdf
- SAS. (2014c). *Industry solutions.* Retrieved from http://www.sas.com/en_us/industry.html
- SAS. (2014d). Pricing and Licensing Information. Retrieved from <https://www.sas.com/order/product.jsp?code=PERSANLBNDL>
- Success stories for SPSS. (2014). Retrieved from http://www-01.ibm.com/software/success/cssdb.nsf/topstoriesFM?OpenForm&Site=spss&cty=en_us
- UCLA: Statistical consulting group. (2014). *What statistical analysis should I use? Statistical analyses using SAS.* Retrieved from <http://www.ats.ucla.edu/stat/sas/whatstat/whatstat.htm>
- Vance, A. (2009). Data analysts captivated by R's power. *The New York Times.* Retrieved from http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all&_r=0

- Why use R. (2014). Retrieved from <http://www.inside-r.org/why-use-r>

Trademarks

ActivStats, DataDesk, ForecastX, Java, JMP, Maple, Mathcad, Mathematica, Matlab, Megastat, Minitab, Python, SAP, SAS, SAS Enterprise Miner, SPSS, Stata, and StatExact are registered trademarks of their respective companies.

Author Biographies

Ceyhun Olgur, CPIM is a professor of information and decision sciences in the College of Business at Valparaiso University. Dr. Olgur has published in Operations Management Research, Decision Sciences Journal of Innovative Education, Quality Management, Production Planning & Control, INTERFACES and OMEGA.

Michelle Kleckner is a graduate student in University of Michigan in Bio-statistics. She earned a BS in mathematics from Valparaiso University.

Yang Li is an undergraduate student in Valparaiso University majoring in accounting.