4-2017

Selection of Statistical Software for Solving Big Data Problems for Teaching

Ceyhun Ozgur
Valparaiso University

Min Dou
Valparaiso University

Yang Li
Valparaiso University

Grace Rogers
Valparaiso University

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

5-1-2017

Selection of Statistical Software for Data Scientists and Teachers

Ceyhun Ozgur
Valparaiso University, ceyhun.ozgur@valpo.edu

Min Dou
Valparaiso University, Min.Dou@valpo.edu

Yang Li
Valparaiso University, yang.li1@valpo.edu

Grace Rogers
Valparaiso University, grace.rogers@valpo.edu

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Selection of Statistical Software for Data Scientists and Teachers

Selection of Statistical Software for Data Scientists and Teachers

Ceyhun Ozgur
Valparaiso University
Valparaiso, IN

Min Dou
Valparaiso University
Valparaiso, IN

Yang Li
Valparaiso University
Valparaiso, IN

Grace Rogers
Valparaiso University
Valparaiso, IN

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 sought by employers and to align this with the software taught by universities. This paper will examine multiple data analysis software – Excel add-ins, SPSS, SAS, Minitab, and R – and it will outline the cost, training, statistical methods/tests/uses, and specific uses within industry for each of these software. It will further explain implications for universities and students.

Keywords: Big data, Excel, R, SAS, SPSS, statistical software, data scientist

Introduction

In the age of big data, technology has transformed how business decisions are made. According the McKinsey Global Institute, "decision making will never be the same; some organizations are already making better decisions by analyzing entire datasets from customers, employees, or even sensors embedded in products" (Manyika et al., 2011, p. 5). In addition to intuition and judgment, business personnel use various software to draw conclusions from data sets and to thereby make decisions.

In measuring the popularity of many data analysis software, Muenchen (2014) noted discovering the software skills that employers are seeking would "require a time consuming content analysis of job descriptions" (para. 17). Muenchen found other ways to determine the statistical software skills that

Dr. Ozgur is a Professor of Information and Decision Sciences in the College of Business. Email him at: ceyhun.ozgur@valpo.edu.

employers seek. One method is to examine which software they currently use. Muenchen cited a survey conducted by Rexer Analytics about the relative popularity of various data analysis software in 2010. The results are pictured in Figure 1. Data miners use R, SAS, and SPSS the most. It can be inferred these are the software skills that the greatest proportion of employers will continue to look for in their potential employees. 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 in their current employees.

Another method used by Muenchen (2014) was to study software skills employers currently seek as they try to fill positions. A rough sketch of statistical software capabilities sought by employers was put together by perusing the job advertising site Indeed.com, a search site the comprises major job boards – Monster, Careerbuilder, Hotjobs, Craigslist – as well as many newspapers, associations, and company websites. The results are summarized in Figure 2.
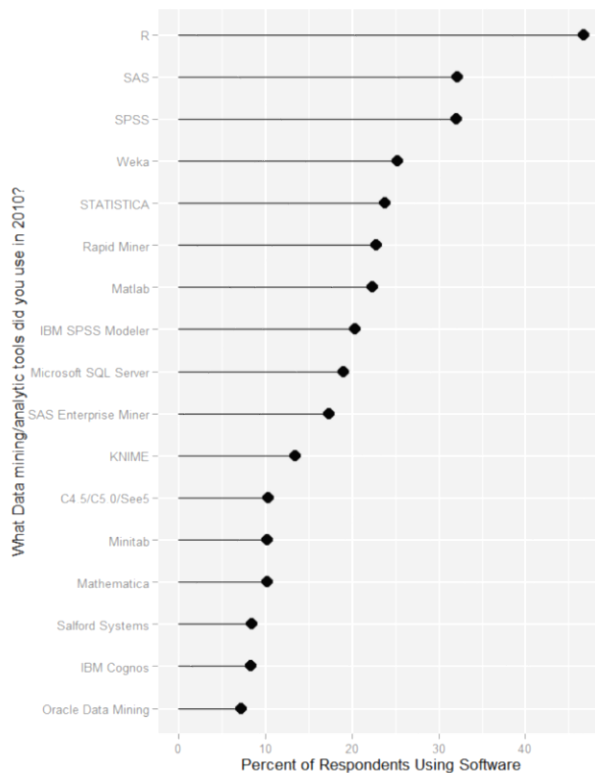

Figure 1. 2010 Rever Analytics survey results of analytic tools (Muenchen, 2014)
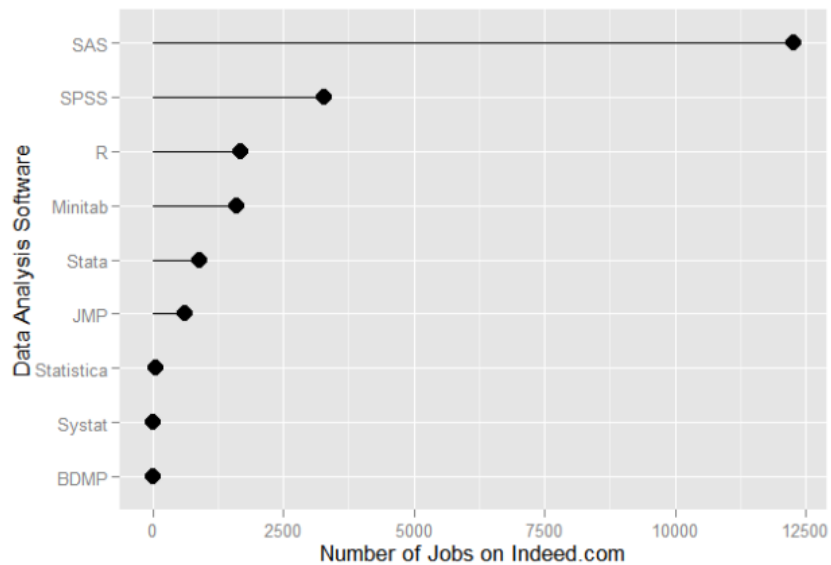
Figure 2. Jobs requiring various software (Muenchen, 2014)

As seen, in contrast to R's greater usage in companies over SAS, illustrated in Figure 1, jobs requiring SAS led to more positions than any other data analysis software. For employers, SPSS and R skills finished in second and third place. This second estimation method of Muenchen (2014) measured the software skill deficits in the job market. It seemed the demand for people with SAS skills outweighs the number of individuals with this capability. One reason for this disconnect could be that college and university faculty are not teaching SAS skills in proportion to the demand for these skills (Lofland & Ottesen, 2013).

To assess this potential disconnect, a non-random survey was conducted with faculty from eighteen departments, which included small and large, state and private, undergraduate and graduate, and East and West, with the results compiled in Table 1. As expected, there was a discrepancy between the software taught and the software sought. SAS led in job openings, but data analysis software taught at those universities did not reflect it. Only a few departments had faculty who taught SAS more than R or SPSS.

The faculty at some departments did not teach any software at all. For example, at Valparaiso University, faculty in the Information and Decision Sciences Department did not teach statistical software, although in certain courses the faculty utilized SPSS, SAS, and R. Excel was the most applicable software used.

Table 1. Results from a survey of statistical software packages taught (Compiled by Kleckner, 2014)

Department	Software Taught at Grad Level	Software Taught at Undergrad Level
Large, Midwestern, State Universities		
Actuarial Science	SAS, Excel, Mathematica	SAS
Mathematics	None	none
Marketing	SAS, SPSS, JMP	N/A
Marketing	SPSS, Excel*	SPSS, Excel*
Large, Southeastern, State Universities		
Statistics	SAS, R, SAS Enterprise Miner	SAS, R, JMP
Engineering	Excel, JMP, Matlab, Mathematica, Mathcad	SAS, Excel, JMP, Matlab, Mathematica, Maple, Mathcad
Economics	N/A	SAS, R, ForecastX, GRETL
Economics	No Graduate Program in Economics	SPSS, Excel, Stata
Information Systems & Decision Sciences	SAS, SPSS, Excel, Megastat, JMP, SAP, Minitab, Matlab, Stata, Mathematica*	SAS, SPSS, Excel, Megastat, JMP, SAP, Minitab, Matlab, Stata, Mathematica*
Medium, Northeastern, Private Universities		
Statistics	SAS, R, Excel, Minitab, JMP, Matlab, Python	N/A
Mathematics	SAS, R, JMP, Matlab, DataDesk, ActivStats*	SAS, R, JMP, Matlab, DataDesk, ActivStats*
Medium, Southeastern, Private University		
Biostatistics	SAS, SPSS, Minitab, Mathematica, Fortran, StatExact, Spatial Stat, C, C++	No Undergraduate Program in Biostatistics
Small, Midwestern, Private Universities		
Mathematics & Computer Science	N/A	SAS, Excel
Mathematics	No Graduate School	SPSS, Excel, Minitab, Mathematica
Statistics	No Graduate School	R
Economics	No Graduate School	Minitab, GRETL
Small, Southern, Private Universities		
Computational and Applied Mathematics	Matlab, C, C++	Matlab, C, C++
Statistics	SAS, SPSS, R, Excel, JMP, Matlab, Mathematica, Stata	JMP, Stata

Note: * These schools did not specify whether the software listed were for graduate or undergraduate students, so we assumed both

This survey was not random, and therefore they cannot be generalized throughout the United States. However, within the sample, there was a trend seen in quantitative, engineering, and business departments, where the use of statistical packages were not aligned to the skills required by employers.

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. Faculty must also consider the cost and time effectiveness of incorporating each software into their curriculum. Further, faculty in specific departments within the school should consider which software best fits their area of study.

Purpose of the Study

The purpose of this study is to gather and condense the necessary information for teaching statistical software. It will assist faculty in their software choices, and it will help their counterparts in business decide which software is best to bring their workforce to the next level of capability. This has increased importance as big data analysis becomes a necessity in business, as Manyika et al. (2011) noted.

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. Big data can no longer be ignored, as noted by the successes of companies where it is invoked as compared to less-modern competitors (Manyika et al., 2011).

Computer software can be written to flexibly support statistical practice (Buchan, 2000). Hence, the focus of this study is on SAS, SPSS, and R software, because both methods in Muenchen's (2014) study indicated they are the three most competitively sought software in business.

Minitab for Teaching Purposes

Minitab's Quality Trainer teaches users how to analyze data online. This multimedia course includes animated lessons that bring statistical concepts to life, and interactive quizzes that give real-time feedback. Hands-on exercises walk the user through applying statistics with Minitab Statistical Software, so knowledge may be put to use immediately.

Quality Trainer contains nine chapters with 141 interactive lessons that can be repeated. It covers statistics needed to analyze quality improvement data, including Basic Statistics, Control Charts, Process Capability, ANOVA, DOE, and more. Easily implementation of projects using a comprehensive collection of more than 100 tools specifically designed for each task. These built-in templates promote greater speed and accuracy.

Below is a list detailing Minitab guide use of statistical and other tools to manage projects:

- **Value Stream Mapping:** Establish the flow of materials and information through your organization. Streamline processes to add value that meets customer expectations.
- Fishbone Diagram: Identify every relevant element of your process and refine the scope of complex projects.
- On-Demand Coaches: Receive the expert guidance you need to complete every step of your project. Add your own instructions or information to any Coach.
- Process Mapping: Construct high-level or detailed flow charts that help you understand and communicate all the activities in a process. Assign variables to each shape and then share them with other tools.
- FMEA (Failure Modes and Effects Analysis): Identify the potential causes for a product or process failure, anticipate the resulting effects, and prioritize the actions needed to mitigate them.
- Pugh Matrix: Compare product design proposals and improvement strategies and determine which ones best fulfill your customer requirements and organizational goals.
- Capture Analysis: Identify and record the important and relevant sections of your Minitab analyses.
- Financial Analysis: Estimate your project savings and the timeframe for realizing them.
- Project Risk Assessment: Evaluate whether a potential project can be successfully completed on time.
- Stakeholder Analysis: Summarize the impact your stakeholders have on your project so you can more effectively leverage their support and address their concerns.
- 5S Audit: Evaluate process conditions relative to 5S best practices and track the ongoing implementation of 5S improvements and controls.

- SIPOC (Supplier-Input-Process-Output-Customer) Analysis: Identify every relevant element of your process and refine the scope of complex projects.
- C&E (Cause and Effects) Matrix: Save time determining what *X* variables to address by comparing and evaluating their potential to impact your goal.
- Y Metrics Chart: Evaluate the progress of your project over time in relation to its baseline and goal.
- Insert Team Members: Easily add team members to your project from your e-mail address book or other file.

Excel Add-Ins

Add-ins are programs that add optional features and commands. With regard to Microsoft Excel, there are add-ins for a multitude of purposes: data analysis, presentation, investment, business, personal, utilities, and productivity tools, and organization. Within data analysis are the Analysis Toolpak, Solver, MegaStat, and PHStat. Both MegaStat and PHStat access codes come with a textbook. However, if an access code isn't available for PHStat, the MegaStat add-ins are available separately from McGraw-Hill (http://highered.mheducation.com/sites/0077425995/information_center_view0/index.html) and Pearson (https://wps.aw.com/phstat/), respectively. Although the Analysis Toolpak and Solver are free add-ins, MegaStat is not.

MegaStat Training

With the current focus on STEM (science, technology, engineering, and mathematics), students and workers may already be familiar with Microsoft Excel or similar spreadsheet software. Building on this familiarity, Burdeane (O. Burdeane, personal communication, January 29, 2014) explained, "Since MegaStat looks and works like Excel, almost anyone could use it to generate some output with just a few minutes of training". MegaStat has dialog and input boxes, buttons, and checkboxes that work largely the same as those in 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 sufficient guidance to effectively use this software.

Statistical Methods/Tests/Uses

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 generate random numbers (McGraw-Hill Education, 2014). SPSS and SAS, for example, have more advanced options, "especially in the area of multivariate statistics" (O. Burdeane, personal communication, January 29, 2014). However, "MegaStat can handle most things encountered by non-PhD statisticians" (O. Burdeane, personal communication, January 29, 2014).

The major caveat for this inexpensive and easy-to-use software is its size capability. For example, Burdeane (O. Burdeane, personal communication, January 29, 2014) experimented with the number of data points that MegaStat can handle, and noted, "I did find a file with 10 columns and 152630 rows. That is over 1.5 million data points and MegaStat did a descriptive statistics analysis on it in about 10 seconds." Although the capability to analyze a million and a half data points sounds quick, this capability may not meet the demand of large companies, because "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" (Troester, 2012, p. 1). Extracting this data and making use of it using MegaStat is not feasible. Other restrictions of MegaStat include its limitation to twelve independent variables in multiple regression and restrictions on variables and table size (O. Burdeane, personal communication, January 29, 2014).

Burdeane (personal communication, January 29, 2014) opined

> 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.

Burdeane also suggested many analyses do not require major packages, like SAS, SPSS, and R, but statisticians stick to them because they are comfortable.

However, personnel in industry 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).

Other Excel Add-Ins

- Analyse (www.analyse-it.com) Standard Edition
- XLStat (www.xlstat.com) from Addinsoft's website.
- NumXL (www.spiderfinancial.com/products/numxl)
- Quantum XL (www.sigmazone.com)

SPSS

SPSS, originally termed Statistical Package for the Social Sciences, was released in 1968 as a software designed for the social sciences. A series of companies subsequently acquired SPSS, ending with International Business Machines (IBM), the current owner, during which time the product's user base was expanded. Therefore, its former acronym was replaced with Statistical Product and Service Solutions to reflect the greater diversity of its clients. Along with Minitab, it is one of the leading statistical packages used in the social and behavioral sciences.

Cost

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 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 twelve 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.

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

distributers, like Creation Engine, On the Hub, StudentDiscounts.com, Studica, or ThinkEDU (IBM, n.d.c). For example, on the Creation Engine website, students can buy the SPSS Statistics Premium GradPack (IBM, n.d.a).

Training

Crossman (n.d.) addressed the difficulty of using SPSS for the first time:

> 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. (SPSS section, para. 1)

Although SPSS does look similar to typical spreadsheet applications like Excel, and its ease of use is very comparable to Excel as well, the cells cannot be manipulated in spreadsheet fashion.

Statistical Methods, Tests, Uses

"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, para. 3). 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, para. 3). SPSS software furthermore has a much more convenient platform for performing statistical tests. For instance, performing a one-sample t-test in Excel (without a plug-in) 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, para. 4). 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.

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, conjoin, direct marketing, exact tests, neural networks, amos, sample power, and

visualization designer, in addition to all of the packages from the professional bundle (IBM, n.d.b). SPSS is also useful for generating plots of distributions and trends, charts, and tabulated reports.

Specific Uses in Industry

On its website, prospective SPSS clients can read about applications in fields like automotive, banking, chemical and petroleum, computer services, consumer products, education, electronics, and energy and utilities. They can also access a list of SPSS's clients. Below are specific examples of SPSS at work within business.

- Infinity Insurance uses SPSS's predictive analytics feature to detect fraudulent claims (IBM, n.d.e).
- "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 fundraising (IBM, n.d.d).
- The Guardia Civil, Spain's very first national law enforcement agency, has investigated crimes and psychology using SPSS (IBM, n.d.d).
- 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" (IBM, n.d.d).

SAS

SAS (Statistical Analysis System) is a commercial statistical package that was developed during the 1960s at North Carolina State University as part of an agricultural research project. Its usage has grown considerably. Ninety-one of the top one hundred companies on the 2013 Fortune Global 500 list use the software (SAS Institute, n.d.a). SAS does not run on Mac computers very easily. One way to run the software on a Mac computer is through parallels, where users buy and run the Windows interface as well.

Cost

An individual license of the Analytics Pro version is available on an annual basis, with a price reduction for subsequent years. With a few more features than the Analytics Pro system, the Visual Data Discovery package is more expensive, although renewals are available with a price break. It is important to consider the added costs if a user wishes to perform data analysis for the benefit of some other party. A different license must be obtained by consulting a SAS representative (SAS Institute, n.d.c).

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 Institute, n.d.c).

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 (Lofland & Ottesen, 2013, p. 3).

In addition to the software license, there is also considerable cost time in the form of installation. Lofland and Ottesen (2013) explained that "SAS can be difficult for users to obtain and the initial installation is sometimes tricky […] long and difficult" (p. 3). However, SAS does not require users to install additional packages.

Training

Crossman (n.d.) claimed, "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" (SAS section, para. 1). 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, p. 1). However, 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. 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; even if users do not have complete understanding of the software, they can run it. Although there exists no easy way to calculate the number of books written about SAS, Muenchen estimates it by searching for books published with SAS in their title and found that close to 500 were published between 2001 and 2011 (as cited in Lofland & Ottesen, 2013). Regarding user support, Lofland and Ottesen (2013) observed

> 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. (p. 3)

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

Statistical Methods, Tests, Uses

SAS's Analytics Pro bundle comes with three of the most popular SAS products: Base SAS, SAS/STAT, and SAS/GRAPH. The 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 code. The list includes thirty-two tests that come from statistical categories such as regression, factor

analysis, discriminant analysis, ANOVA, non-parametric tests, and correlation (University of California, Los Angeles [UCLA]: Statistical Consulting Group, n.d.). The full list can be seen in the Appendix below.

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" (http://support.sas.com/software/) has a list of these programs.

Also on this page, SAS has additional packages to access that are industry-specific. For example, there is an SAS Drug Development package that "enables the efficient development, execution and management of analysis and reporting activities for clinical research," (SAS, n.d.b) an 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, payments and nonmonetary transactions," (SAS, n.d.b) and an SAS Risk Management for Insurance package that "implements the Solvency II standard model approach for calculating risk-based capital with [its] comprehensive solution for preforming risk analysis and risk-based capital calculations" (SAS Institute, n.d.b). In addition to 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 of disorganized data. Real life data is 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 (Lofland & Ottesen, 2013, p. 3-4). 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 Lofland and Ottesen (2013), who explained, "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" (p. 3-4).

Specific Uses in Industry

SAS has built-in, functional packages for many specific industries, including health-care, banking, insurance, law enforcement, communications, retail, casinos, utilities, sports, and more. To follow 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

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 more than two million users according to an R Community website (Revolution Analytics, n.d.a).

Cost

R is free and is downloadable from the Internet, with 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. Although the base for R is very easy to install, users must download packages to perform specific analyses, which can be very time-consuming (Lofland & Ottesen, 2013, p. 3-4). For example, currently there are 5,508 available packages, and this number grows weekly if not daily (Comprehensive R Archive Network [CRAN], n.d.). This provides many options, but searching through the assemblage of choices can be difficult and time-consuming.

Training

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. In many ways, though, R is comparable to other programming languages. For instance, similar to many other languages, it is a command line interface. Additionally, 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 will often speed up the learning process for trainees since 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, Pregibon claimed R "allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems (Vance, 2009, para. 4). Vance (2009) also noted, "R has quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use" (para. 3). R is not as daunting as other languages, having very natural and expressive syntax for data analysis. 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 (R Core Team, 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 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 in order to use the software. Lofland and Ottesen (2013) stated, "[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" (p. 3). Although R may be useable 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.

Lofland and Ottesen (2013) also explained 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" (p. 3).

Statistical Methods, 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.)" (Revolution Analytics, n.d.b). Programmers have designed many more packages than just these, including packages for Bayesian statistics, time series analysis, simulation based analysis, spatial statistics, survival analysis, and many, many more (CRAN, 2014). A complete list of packages already designed for R can be found on the R packages website (http://cran.us.r-project.org/web/packages/).

The key feature of R that differentiates it from other statistical software is its acceptance of customization. The aforementioned software have "data-in-data-out black-box procedures" (Revolution Analytics, n.d.b). 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. A "decomp" command, or something of the sort, is all that is needed, and the statistical package will perform the decomposition for them. For example multiplicative decomposition is the forecasted value (F) = Trend × Seasonal × Cyclical × Irregular. 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 pre-packaged functions in the other statistical software, this is not as easy.

R is known for generating appealing charts and tables. The custom charting capabilities of R create "stunning infographics seen" (Revolution Analytics, n.d.a). However, it cannot manage messy data as easily as other available statistical software. Lofland and Ottesen (2013) warned, "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" (p. 3). But, once the data is organized, R is a valuable data analysis performer and graphics creator.

Specific Uses in Industry

The usage of R is diverse in business. Some examples follow.

- Google "taps R for help understanding trends in ad pricing and for illuminating patterns in the search data it collects" (Vance, 2009, para. 24).
- 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 paper, 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 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, the purpose of this study was to investigate their features.

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. Faculty at colleges and universities could improve job placement by preparing students in the specific software that hiring companies use. It is difficult for new data analysts to see the forest for the trees when choosing a statistical programming language (DataCamp Team, 2014).

Students can add a software taught category to their list of traits sought in higher education in order 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 that 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.

References

Boudreaux, D. (2003, March-April). *Java syntax for SAS programmers*. Paper presented to the SAS Users Group International, Seattle, WA.

Buchan, I. E. (2000). *The development of a statistical computer software resource for medical research* (Doctoral dissertation). University of Liverpool, Liverpool, England.

Comprehensive R Archive Network. (n.d.). *Contributed packages*. Retrieved from http://cran.us.r-project.org/web/packages/

Crossman, A. (n.d.). *Analyzing quantitative data: Statistical software programs for use with quantitative data* [Web log post]. Retrieved from http://sociology.about.com/od/Research-Tools/a/Computer-programs-quantitative-data.htm

DataCamp Team. (2014, June 3). What is the best statistical programming language? [Web log post]. Retrieved from https://www.datacamp.com/community/tutorials/statistical-language-wars-the-infograph#gs.eO=sntU

International Business Machines. (n.d.a). *IBM SPSS Statistics Premium GradPack 22*. Retrieved from

http://www.creationengine.com/html/p.lasso?l=SPSS%20Statistics%20Premium%20GradPack&p=18830

International Business Machines. (n.d.b). *SPSS Statistics*. Retrieved from http://www-01.ibm.com/software/analytics/spss/products/statistics/buy-now.html

International Business Machines. (n.d.c). *SPSS Statistics GradPack*. Retrieved from http://www-03.ibm.com/software/products/en/spss-stats-gradpack/

International Business Machines. (n.d.d). *Success stories for SPSS*. Retrieved from http://www-01.ibm.com/software/success/cssdb.nsf/topstoriesFM?OpenForm&Site=spss&cty=en_us

International Business Machines. (n.d.e). *Why SPSS software?* Retrieved from http://www-01.ibm.com/software/analytics/spss/

Lofland, C. L., & Ottesen, R. (2013, April-May). *The SAS versus R debate in industry and academia*. Paper presented to the SAS Global Forum 2013, San Francisco, CA.

Manyika, J., Chi, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

McGraw-Hill Education. (2014). *MegaStat*. Retrieved from http://glencoe.mcgraw-hill.com/sites/0010126585/student_view0/megastat.html

Muenchen, R. A. (2014). *The popularity of data analysis software*. Retrieved from http://r4stats.com/articles/popularity/

R Core Team. (2014). *An introduction to R*. Retrieved from http://cran.r-project.org/doc/manuals/R-intro.html#Statistical-models-in-R

Revolution Analytics. (n.d.a). *What is R?* http://www.inside-r.org/what-is-r

Revolution Analytics. (n.d.b). *Why use R?* http://www.inside-r.org/why-use-r

Robbins, S. (2012, June 7). *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 Institute. (n.d.a). *About SAS*. http://www.sas.com/en_us/company-information.html

SAS Institute. (n.d.b). *Industry solutions*. Retrieved from http://www.sas.com/en_us/industry.html

772

SAS Institute. (n.d.c). *Pricing and licensing information*. Retrieved from https://www.sas.com/order/product.jsp?code=PERSANLBNDL

Troester, M. (2012). *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/whitepaper/wp_46345.pdf

University of California, Los Angeles: Statistical Consulting Group. (n.d.). *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, January 6). 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

Appendix A: 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
- Non-parametric correlation
- Simple logistic regression
- Multiple regression
- Analysis of covariance
- Multiple logistic regression
- Discriminant analysis
- One-way MANOVA
- Multivariate multiple regression
- Canonical correlation
- Factor analysis

(UCLA: Statistical Consulting Group, n.d.)