

Statistical Software Programs used for Business Research



DOI: 10.26573/2021.15.3.5
Volume 15, Number 3
September 2021, pp. 195-206

Ceyhun Ozgur
Sanjeev Jha

Information and Decision Sciences
Valparaiso University, USA

Yiming Shen

Rochester Institute of Technology, USA

Ali Arminian

Ilam University, Iran

In this paper we discuss software programs used for teaching business courses and used in business research. The software programs used are MATLAB, R Studio, Microsoft Excel, SPSS, SAS and Python. This paper goes into details about the functions of each software program and how each program is proficient and programmed for different areas of research. Using tables and figures, we discuss the ease of use per program, the difficulty based on a scale for beginner uses, the storage capacity of each programming language, the availability to business students and researchers and how applicable each program is with a basic knowledge of how to use it. The cost of each statically programming package is also covered. The detailed functions of the various programs are covered in figures and tables in this paper. It is a well-known fact that MATLAB, R, Python, Excel, SPSS and SAS are the most important five languages to be learned for data analysis.

Keywords: Big Data, Teaching R, Teaching Python, Teaching SPSS, Teaching SAS, Teaching Microsoft Excel, Teaching MATLAB, Demonstrations of Examples for Teaching Statistical Software Packages

1. Introduction

MATLAB, R, Excel, SAS, SPSS and Python are analytical software programs which configure statistical analyses by producing outputs in the form of graphs. MATLAB is a matrix-based program that expresses math computations for engineers and science-based field research. R is open source. It is a versatile statistical software program that can be used in a range of changing field preferences. It is highly standardized. SAS is a paid software system that provides high performance analytics operations. Through using this software program organizations can identify and investigate the life cycle of any inquiry. Excel is a free program provided by Microsoft, available to everyone that has Windows. SPSS is a statistical analysis software program provided by IBM. Python is a programming language created by Guido van Rossum which was designed to emphasize code readability. When it comes to data science one of the most common points of debate is R vs SAS vs Python vs Excel vs SPSS vs MATLAB. It is a well-known fact that MATLAB, R, Python, Excel, SPSS and SAS are the most important five languages to be learned for data analysis.

Amongst all six statistical software programs, each have different storage capabilities. For example, R Studio requires 250 GB SAS requires 10-15 GB to

download. While Python requires 2 GB and Excel only requires 1 GB of RAM as the minimum system requirement for the 32-bit version of Windows 10. SPSS rule of thumb is to have four times as much space. 2 GB hard drive space and 4 GB of RAM will be used. Booth and Ozgur discuss how predictive modeling are utilized in evaluation of technical acquisition performance using survival analysis. (Booth, Ozgur 2019).

Table 1 Comparison of Software Programming Languages

	SPSS	SAS	R Studio	Python	MATLAB	Excel
Ease of use 1-6 (1 being the easiest)	4	6	5	3	2	1
Difficulty	Moderately Difficult	Most Difficult	Not difficult if user knows of package plan assistance.	Moderately Difficult	Not Difficult if codes are known	Not Difficult
Storage	2 GB of Hard Drive space, 4 GB of RAM	Requires 10-15 Gb to download	250 GB. R will not need as much scratch space as SAS.	2 Gb	Min: 3.5 GB. Typically, 5-8 Gb used.	1GB of RAM is the min system requirement for the 32-bit version of Windows 10
Availability to students and researchers 1-5 (1 being the least prevalent)	2 moderate)	1	2 (moderate) depending on if package programs are used.	2 (moderate) users must know software language to operate.	2 (moderate) users must know software language to operate.	3
Cost	\$99.00 / month subscription	\$8,000/year	Free Download Available	\$0-\$8.00/month /user	\$95.00 /home user	Free with the purchase of Microsoft
Application	2 nd Most Applicable.	Most Applicable.	Very Applicable with knowledge of package programs.	Applicability is to be determined by user.	Applicability is to be determined by user.	Only Applicable for small to medium sized problems.

In the above table we compare SPSS, SAS, R Studio, Python, MATLAB, and Excel in terms of the program's ease of use, difficulty, storage capacity, availability to students and researchers, overall cost, and each programs applicability in the real world.

2. MATLAB

MATLAB is an integrated software program that numerically computes a high level statically language. As well as visualization in the form of graphics and simulations which can be used for data analysis exploration. Application of this software program can help the user develop models and algorithms in a system interface.

R Studio

R Studio is a programming language used in different fields. R is an open source and easy to access, supported by the R Foundation, which is a statistical computing foundation. R is widely used by data miners for developing data analysis with the aid of pre-packaged programs. These pre-packaged have a command line interface and several graphical front-ends which are available.

Microsoft Excel

Microsoft Excel is more than a spreadsheet as it is capable of doing mathematical analysis for the user. Given it can encode data, create complex graphs, and manipulate numbers through formula functions. Provides easy reference to input. All while being able to manipulate numbers in a mathematical environment.

SPSS/ Statistical Package for the Social Sciences

SPSS analyses data to solve research problems through an interface that is easy to use. It has the capabilities of advanced statistical procedures. These procedures can use extensions such as R, Python which ensure accurate data analysis and progressive decision making. This programming language follows a spread sheet format similar to Excel. Users are able to solve large scale problems.

SAS/ Statistical Analysis System

SAS is a programming language that uses a common spread sheet layout which results in different statistical analyses in the form of tables, graphs, RTF, PDF, HTML documents. SAS is an expensive software language however schools and business can afford it. SAS is involved in many different sectors of business. Helping users make cost effective decisions.

Python

Python is a programming language designed to emphasize code readability. Python has variants for C and Java programming languages. The C variant is known as Python and is designed to give Python the advantages of C. One of these characteristics is in terms of performance. The variant can act both as an interpreter and at the same time as a compiler. Python has a wide range of applications.

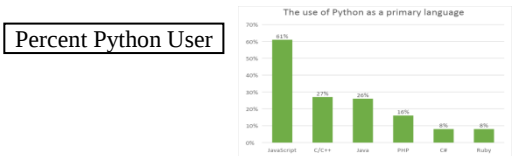


Figure 1 Use of Python as a Primary Language
(Python Developers Survey 2016: Findings)

In this figure we describe how many users (JavaScript, C/CC+, Java, PHP, C#, Ruby) prefer Python as their primary language.

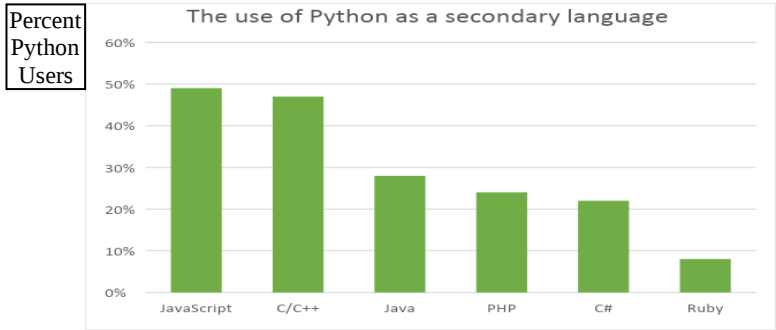


Figure 2 Use of Python as a Secondary Language
(Python Developers Survey 2016: Findings)

In Figure 2, about 49% of user’s (JavaScript, C/CC+, Java, PHP, C#, Ruby) prefer to use python as their secondary development language.

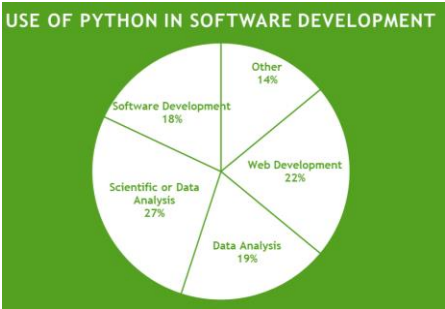


Figure 3 Use of Python in the Software Development
(Python Developers Survey 2016: Findings)

There are about 46% of developers which use Python programming language as their Data analysis tool instead of traditional programmers or Web developers.

In an early paper Ozgur, et.al. In journal of Data Science compared the effectiveness of MATLAB, Python and R in a teaching environment. For example, Python is the programming language that is based on C which contains a standard library which is structured to focus on different modules for threading, networking and databases.

Table 2 Python vs. R in Google Citations

	Google result-R	Google result-python	Ratio of R / python
XX + programming	2,430,000	72,700,000	0.33
XX + data collection	4,200,000	1,520,000	2.76
XX + statistics	108,000,000	29,000,000	3.72
XX + model	154,000,000	8,270,000	18.62
XX + code	130,000,000	2,710,000	47.97
	79,726,000	10,375,000	15

In this table, using Google citations we can see the prevalence of a Python citation versus an R Studio web search.

Table 3 Comparisons of R vs. Python Books and Papers

	Books-R	Books-python	Ratio of R / python	Papers-R	Papers-Python	Ratio of R / python
programming for statistics	207	13	15.92	2,680,000	47,600	56.3
Data	6970	310	22.48	9,290,000	205,000	45.32
statistics	4235	42	100	5,470,000	104,000	52.59
model	8631	126	68.5	7,830,000	163,000	48.03
code	6084	105	57.94	4,650,000	158,000	29.43
	5225.4	119.2	52.968	5984000	135520	46.334

In this table, we compare the books for R versus books for Python. First, we compare the scholarly papers for R versus Python. We show the ratio of R books/ papers versus Python books/ papers.

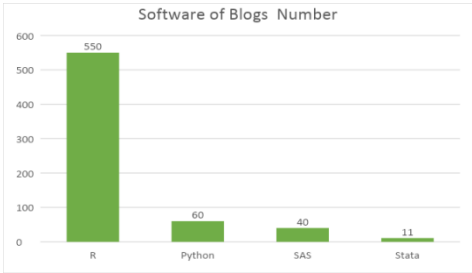


Figure 4 Software Blogs: Number of Blogs Devoted to each Software Package on April 7, 2014, and the Source of the Data

R’s blogs have an impressive number of 550. For Python, only 60 blogs that were devoted to the SciPy subroutine library were found. SAS 40 blogs was an impressive figure given that Stata only possessed 11 blogs.

While searching for a list of blogs related to software, individual blogs were found which related to software. Unfortunately, the list was not kept updated, and would be far too time consuming to deal with.

If you know of other lists of relevant blogs, please inform us. They will be added to the list.

Internet blogs are written by passionate people who speak about problem solving methods and software. Blogs contain information that has the potential to sway the popularity of a software packages.

```
1 import time
2 timer_start = time.time()
3
4 for i in xrange(1, 100000):
5     print "Hi!"
6 #End for
7
8 print "Finished!"
9
10 timer_end = time.time() - timer_start
11
12 print timer_end
```

```
1 ptm <- proc.time()
2
3 for(i in 1:100000) {
4     print("Hi")
5 }
6
7 print("Finished!")
8
9 print(proc.time() - ptm)
```

Figure 5 Computer Codes for R and Python

In this illustration of computer codes for R and Python we can see how both programming Python and R are utilized. The top depiction shows simple Python format. While bottom shows an R Studio format. Python: 0.769 seconds / R: 4.86 seconds.

Why Python is Great for Data Science

- Python was released in 1989. It has been around for a long time, and it has object-oriented programming baked in.
- IPython / Jupyter's notebook IDE is excellent.
- There's a large ecosystem. For example, Scikit-Learn's page receives 150,000 - 160,000 unique visitors per month.
- There's Anaconda from Continuum Analytics, making package management very easy.
- The Pandas library makes it simple to work with data frames and time series data.

Figure 6 Python for Data Science

Figure 6 goes into detail about the history and competition of Python programming language.

Why R is Great for Data Science

- R was created in 1992, after Python, and was therefore able to learn from Python's lessons.
- Rcpp makes it very easy to extend R with C++.
- RStudio is a mature and excellent IDE.
- (Our note) CRAN is a candyland filled with machine learning algorithms and statistical tools.
- (Our note) The Caret package makes it easy to use different algorithms from 1 single interface, much like what Scikit-Learn has done for Python

Figure 7 R Studio for Data Science

Figure 7 shows how R can be utilized in Data Science. It also goes into detail about the history and competition of R Studio programming language.

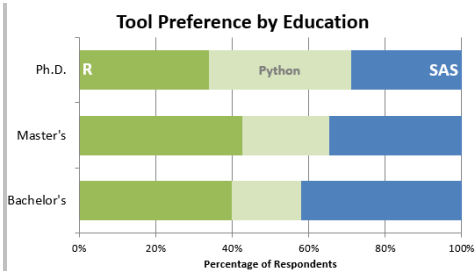


Figure 8 Preference by Education
(<http://www.burtchworks.com/2017/06/19/2017-sas-r-python-flash-survey-results/>)

Figure 8 shows the preference of R, Python or SAS by education levels. At least 40% of each educational level prefers the R Studio software. About 23% of the overall education level of users prefer Python. Roughly 37% prefer SAS.

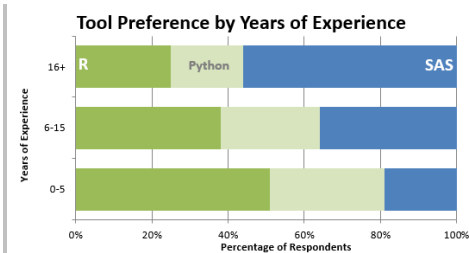


Figure 9 Depicts Tool Preference by the Years of Experience
(<http://www.burtchworks.com/2017/06/19/2017-sas-r-python-flash-survey-results/>)

In Figure 9, about 25% of users who have 16 years or more of experience prefer R. While About 20% of users who have 16 years or more of experience prefer Python and about 55% of users with 16 or more years of experience prefer SAS. About 35% of users who have 6-15 years prefer R, roughly 27% of users with 6-15 years of experience prefer Python and about 38% of users with 6-15 years prefer SAS. Lastly, 50% of users with 0-5 years of experience prefer R, about 32% of users with 0-5 years' experience prefer R, and about 18% of users with 0-5 years' experience prefer SAS.

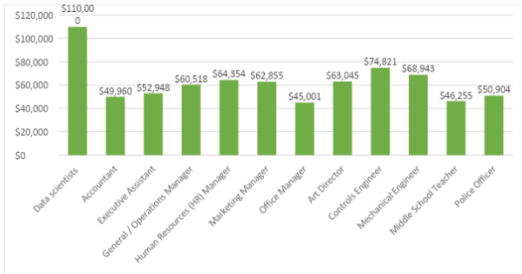


Figure 10 Comparison of Salaries of Different Disciplines against Data Scientists (Data Analysts)

This figure compares the salaries for different occupations. Upon observation of this graph, we can see that Data Scientists earn the highest salary when compared to the other occupations.

In Figure 11 we can see the employers searching for qualified graduates for various positions. The three software languages (R, SAS, and Python) have different growth rates. Rs' growth rate has stayed stable but has increased in 2017. While SAS has also stayed stable over the past two during the fiscal year of 2016-2017 there was an increase to 20%, but sadly decreased to 15% during the 2017-2018 fiscal year. In terms of Python there has been a steady growth rate from 5% in 2014 to approximately 13% in 2017.

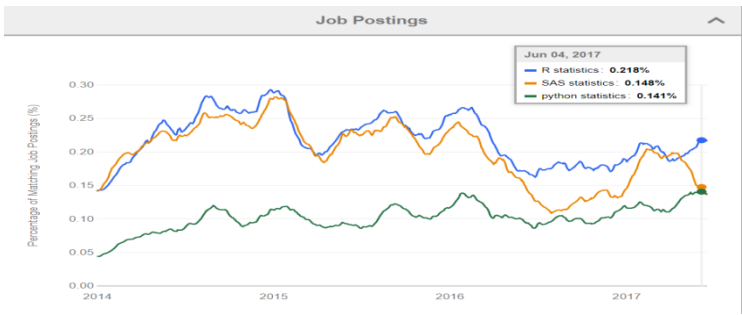


Figure 11 Job Posting of Statistical Software Packages from 2014 to 2017
(<https://www.indeed.com/jobtrends/q-R-statistics-q-SAS-statistics-q-python-statistics.html>)

In Figure 11 shown above, we see the number of graduates interested in working for the three software language companies (R, SAS, and Python). The interest to work for these software companies has had a large fluctuation. For example, SAS in 2014 had a fluctuating increase of interest until 2016 when the interest level dropped significantly. Since the drop the interest level has stayed constant. R stats had a similar trend in fluctuating interest of graduates as SAS, and in 2017 the interest level of both software packages was approximately the same. Finally, for Python in 2014 till 2016 there was a steady growth rate for interested graduates. During the fiscal year of 2016-2017 there was a large spike in interest, but then had a slow decrease.

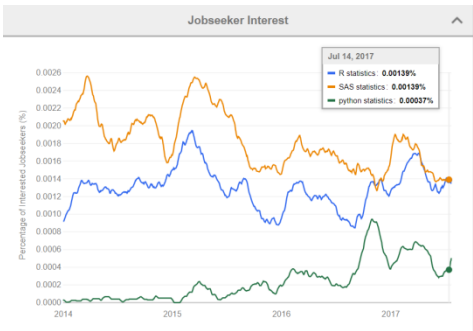


Figure 12 Jobseeker Interested in Statistical Software Packages from 2014 to 2017
(<https://www.indeed.com/jobtrends/q-R-statistics-q-SAS-statistics-q-python-statistics.html>)

Table 4 Recommend for First Learning Statistical Software Language

Language Recommendation Select	Count	Percent
Python	6941	63.11147481
R	2643	24.03164212
SQL	385	3.50063648
C/C++/C#	307	2.79141662
MATLAB	238	2.16402982
Java	138	1.2547736
Scala	94	0.85470085
SAS	88	0.80014548
Other	85	0.77286779
Julia	30	0.27277687
Stata	28	0.25459174
Haskell	17	0.15457356
F#	4	0.03637025
Total	10998	1

In this table, we compare the recommended selection from the users about software programs.

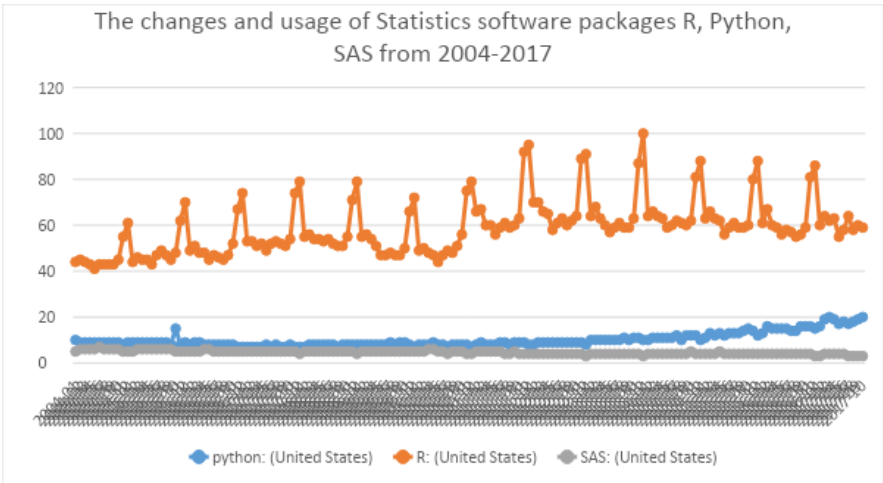


Figure 13 The use of Software Packages and how they have changed over the Years (<https://trends.google.com/trends/explore?date=all&geo=US&q=python, R, SAS>)

This figure shows how the usage of Python, R, and SAS have changed in the United States over the years from 2004-2017 though a linear depiction.

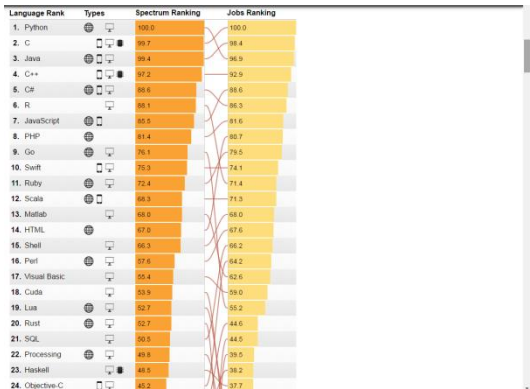


Figure 14 Spectrum and Jobs Ranking of Software Languages in 2017
(https://spectrum.ieee.org/ns/IEEE_TPL_2017/comparison/2017/2016/1/1/1/1/50/1/50/1/50/1/30/1/30/1/30/1/20/1/20/1/5/1/5/1/20/1/100/true/1/25/1/25/1/50/1/25/1/25/1/100/1/100/1/25/1/40/)

This figure addresses the use of R and Python programming Languages for Jobs. The IEEE (Institute Electrical and Electronics Engineers). Spectrum Ranking is a site that will combine 12 metrics from 10 different sites. Some measures that are presented are popularity of job sites/search engines. While at the same time the site can show how much new programming code has been added to GitHub over last year.

Data Bases Used in Software Selection

- Azure SQL
- MySQL
- Oracle

Azure SQL Database is a database as a service. It is an intelligent and provides the user a managed service. A database as a service is a service that provides users with some form of access to a database on-demand. This type of service can be used without the need for setting up any physical hardware base or installing software. All of the administrative tasks and maintenance are taken care of by the service provider, Microsoft. Being part of the Azure cloud means the Azure SQL database comes with some very useful features. It is a single database resource type that creates a database in Azure SQL Database with its own set of resources and is managed via a server. With a single database, each database is isolated. Each Database its own service tier. MySQL is widely used as it is compatible and easy to set up. It's compatible with all major Operating System platforms and all major programming languages. This Database is flexible. It is governed by the GNU (General Public License). MySQL uses this to control how its software is managed.

Both the Azure SQL Database and MySQL deal with relational databases and use SQL as their query language. The architectures of these two systems are different. MySQL is a Database Management System. This means it is a piece of software designed to define, retrieve and manage data in a database. This is different from Azure SQL for that it is not a database on its own. MySQL controls how you store and manipulate data in a database, it is a service. However, its open source nature allows a variety of integration options. MySQL DBMS software can be freely downloaded

from the internet. This service can be modified to suit your system's technology without losing much in terms of performance.

Oracle Database offers its users a cost-optimized and high-performance version of a database that has multi-model database management system, as well as in-memory which enables users to simplify management workloads.

3. LinkedIn Marketing Tool

Use of LinkedIn as a marketing tool for users who use statistical software. LinkedIn is the leading marketing tool when it comes to networks that are aimed for professional use. Over time it has become a highly valued instrument used by recruiters, especially recruiters in the HR business world. The elements of a LinkedIn profile are such as design, methodology and approach. These help the those in the hiring process identify who is the best candidate. There is a feature called LinkedIn Salary product, it was launched in 2016 with the goal of providing insights on compensation distribution to job seekers. This tool is helpful for those trying to find a job using business statistical software for that they can make more informed decisions when discovering job opportunities.

4. References

1. Adrian Kingsley-Hughes for Hardware 2.0 | July 5, 2019 -- 11:54 GMT (04:54 PDT) | Topic: Hardware <https://www.zdnet.com/article/how-much-ram-does-your-windows-10-pc-need-2019-edition/>
2. Booth, D., Ozgur C. The use of Predictive Modeling in the Evaluation of Technical Acquisition Performance using Survival Analysis. Volume 17, Number 3, July 2019.
3. (<https://www.indeed.com/jobtrends/q-R-statistics-q-SAS-statistics-q-python-statistics.html>)
4. (<http://www.burtchworks.com/2017/06/19/2017-sas-r-python-flash-survey-results/>)
5. <http://www.burtchworks.com/2017/06/2017-0-SAS-r-python-flash-survey-results/>)
6. <https://trends.google.com/trends/explore?date=all&geo=US&q=python,R,SAS>.
7. <https://www.quora.com/Is-Python-faster-than-R>
8. <https://www.indeed.com/jobtrends/q-R-statistics-q-SAS-statistics-q-python-statistics.html>
9. <https://www.indeed.com/jobtrends/q-R-statistics-q-SAS-statistics-q-python-statistics.html>
10. Jeremy, H. 2020 Dream Factory Inc.
11. Ozgur, C., Alam, P., and Booth, D., Software Languages for Analytics in Research. 50th Annual Conference of the Decision Sciences Institute.
12. Ozgur, C. S. Jha, Y. Shen. Comparison and Contrast of Statistics Software Packages including R and Python for Teaching Purposes.
13. Ozgur, C. T. Colliau, G. Rodgers, Z. Hughes, E. B. Myer-Tyson. MatLab vs. Python vs. R. Journal of data Science (2017, 355-372
14. PyCharm (January 2017), Python Developers Survey 2016: Findings, <https://www.jetbrains.com/pycharm/python-developers-survey-2016>

15. Pinho G., Arantes J., Marques T., Branco F., Au-Yong-Oliveira M. (2019) the Use of LinkedIn for ICT Recruitment. In: Rocha Á., Adeli H., Reis L., Costanzo S. (eds) *New Knowledge in Information Systems and Technologies. WorldCIST'19 2019. Advances in Intelligent Systems and Computing*, vol 930. Springer, Cham. https://doi.org/10.1007/978-3-030-16181-1_16
16. SAS Institute Inc., *System Requirements for SAS® 9.3 Foundation for AIX®*, Cary, NC: SAS Institute Inc., 2012.
17. *System Requirements for MATLAB R2020B, 1994-2020 THE MATH WORKS, INC.* <https://www.mathworks.com/support/requirements/matlab-system-requirements.html>
18. Xi Chen, Yiqun Liu, Liang Zhang, and Krishnaram Kenthapadi. 2018. How LinkedIn Economic Graph Bonds Information and Product: Applications in LinkedIn Salary. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 120–129. DOI: <https://doi.org/10.1145/3219819.3219921>
19. Zide, Julie; Elman, Ben; Shahani-Denning, Comila. *Employee Relations*, Volume 36, Number 5, 2014, pp. 583-604(22). Emerald Group Publishing Limited.
20. 2001-2020. *Foundation Legal Privacy Policy Powered by Heroku* <https://www.python.org/psf/sponsorship/sponsors/>
21. 2000-2020 R-Tools Technology Inc. https://www.r-studio.com/Unformat_Help/memory_usage.html
22. *Data Source Tool*: <https://trends.google.com> <https://www.indeed.com/jobtrends>

About Our Authors

Ceyhun Ozgur is a professor of Information/Decision sciences at Valparaiso University. He holds a PhD Operations Management. He has published 45 refereed journal articles and presented over 30 papers.

Sanjeev Jha is an associate professor of Information & Decision Sciences in the College of Business at Valparaiso University. He earned his Ph.D. from the University of Illinois at Chicago and had been an assistant professor of decision sciences at the University of New Hampshire prior to coming to Valparaiso University in fall 2015. He teaches data analytics related courses at Valparaiso University and his research interests are in the areas of healthcare analytics. He can be reached at sanjeev.jha@valpo.edu.

Yiming Shen is a third-year PhD candidate at the University of Massachusetts Boston, College of Science and Mathematics of Computational Sciences. He received a bachelor's degree in engineering from Chongqing University and a master's degree in analytics and modeling from Valparaiso University. He is interested in Numerical Optimization and Machine Learning.

Ali Arminian is an assistant professor of plant breeding (biometrical approaches), at Ilam University, Iran.