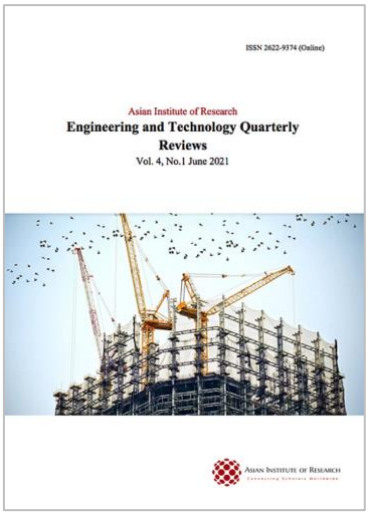

Engineering and Technology Quarterly Reviews

Ozgur, Ceyhun, Jha, Sanjeev, and Shen, Yiming. (2021), Statistical Software Programs Used for Business Research. In: *Engineering and Technology Quarterly Reviews*, Vol.4, No.1, 38-47.

ISSN 2622-9374

The online version of this article can be found at:
https://www.asianinstituteofresearch.org/

Published by:
The Asian Institute of Research

The Asian Institute of Research *Engineering and Technology Quarterly Reviews* is a peer-reviewed International Journal. The journal covers scholarly articles in the fields of Engineering and Technology, including (but not limited to) Civil Engineering, Informatics Engineering, Environmental Engineering, Mechanical Engineering, Industrial Engineering, Marine Engineering, Electrical Engineering, Architectural Engineering, Geological Engineering, Mining Engineering, Bioelectronics, Robotics and Automation, Software Engineering, and Technology. As the journal is Open Access, it ensures high visibility and the increase of citations for all research articles published. The *Engineering and Technology Quarterly Reviews* aims to facilitate scholarly work on recent theoretical and practical aspects of Education.

ASIAN INSTITUTE OF RESEARCH
Connecting Scholars Worldwide

The Asian Institute of Research
Engineering and Technology Quarterly Reviews
Vol.4, No.1, 2021: 38-47
ISSN 2622-9374

Statistical Software Programs Used for Business Research

Ceyhun Ozgur[1], Sanjeev Jha[2], Yiming Shen[3]

[1] Research Professor of Information and Decision Sciences
[2] Associate Professor of Information and Decision Sciences
[3] MS student at Rochester Institute of Technology

Abstract
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 programing package is also covered. The detailed functions of the various programs are covered in Figures throughout and tables through the paper as well. 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 Microsoft Excel, Teaching SAS, Teaching MATLAB, Demonstrations of Examples for Teaching Statistical Software Packages

Introduction

MATLAB, R, Excel, SAS, SPSS and Python are analytical software programs that configure statistical analyses by producing outputs ibn 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 emphasis 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 has 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 4Gb of RAM will be used. Booth and Ozgur discuss how predictive modeling is 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		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	Minimum: 3.5 **GB. Typically, 5-8 Gb used.**	1GB of RAM is the minimum 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	2nd 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 this 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.

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 for that 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 programing 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.

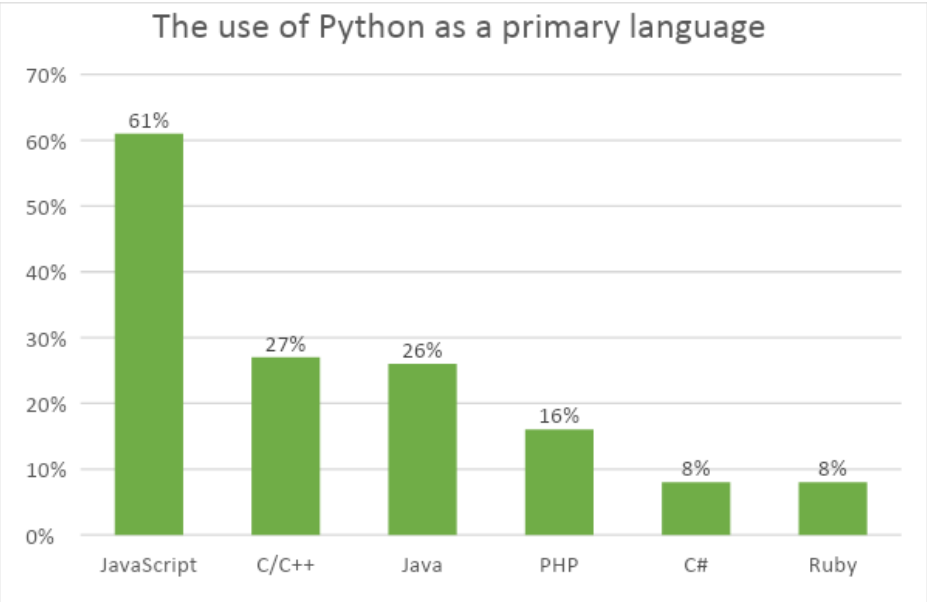

Figure 1: The use of Python as a primary language

(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.

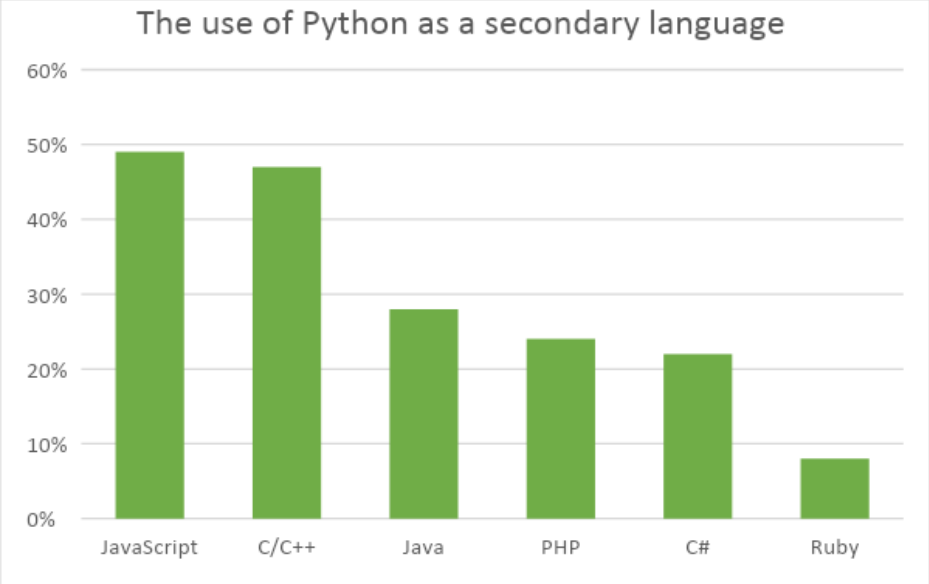

Figure 2: The use of Python as a secondary language.
(Python Developers Survey 2016: Findings)

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

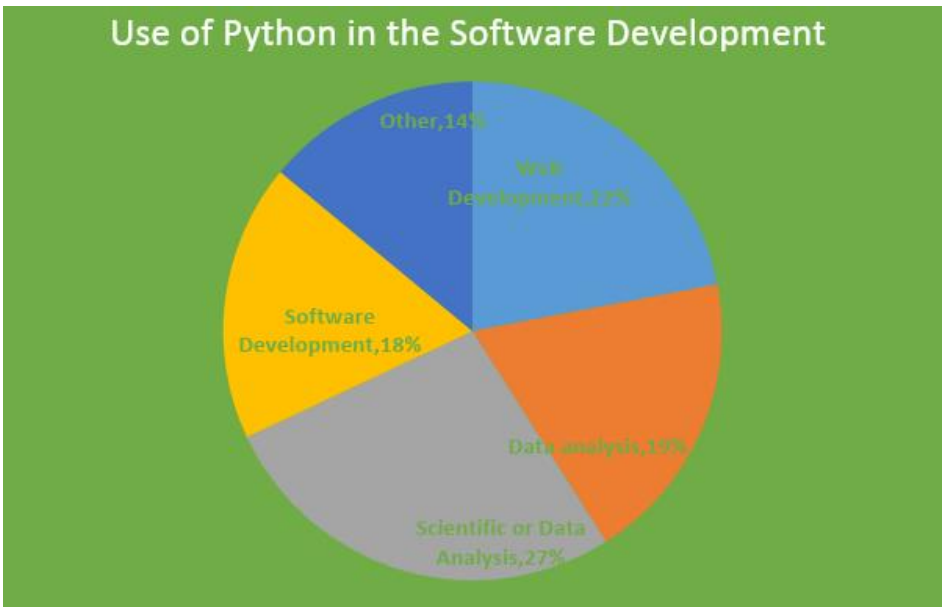

Figure 3: Use of Python in the Software Development.
(Python Developers Survey 2016: Findings)

There are about 46% of developers which use Python programing 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+programing	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.

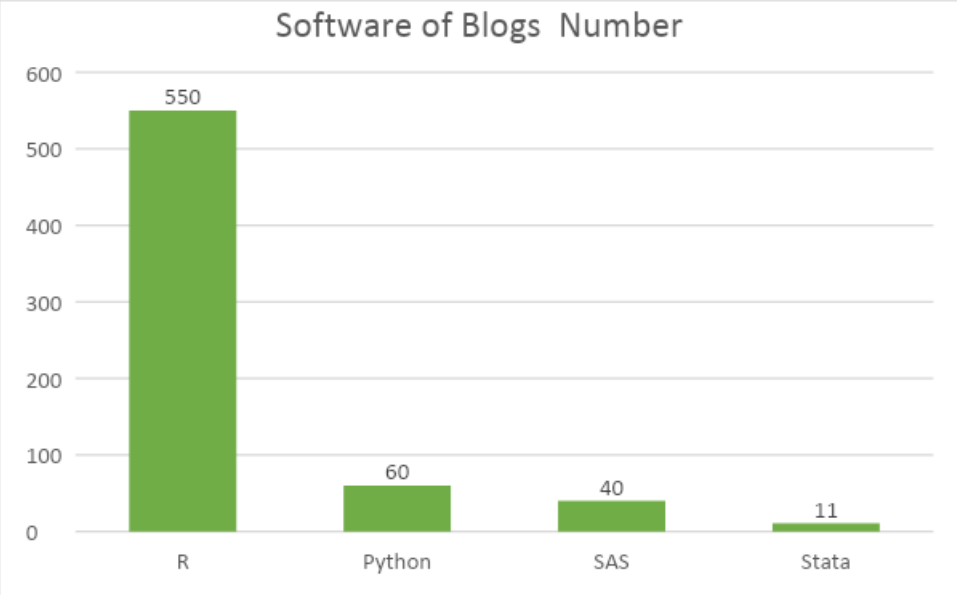

Figure 4: Software Blogs: Number of blogs devoted to each software package on April 7, 2014, and the source of the data.

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.

```python
import time
timer_start = time.time()

for i in xrange(1, 100000):
    print "Hi!"
#End for

print "Finished!"

timer_end = time.time() - timer_st

print timer_end
```

```r
ptm <- proc.time()

for(i in 1:100000) {
  print("Hi")
}

print("Finished!")

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.

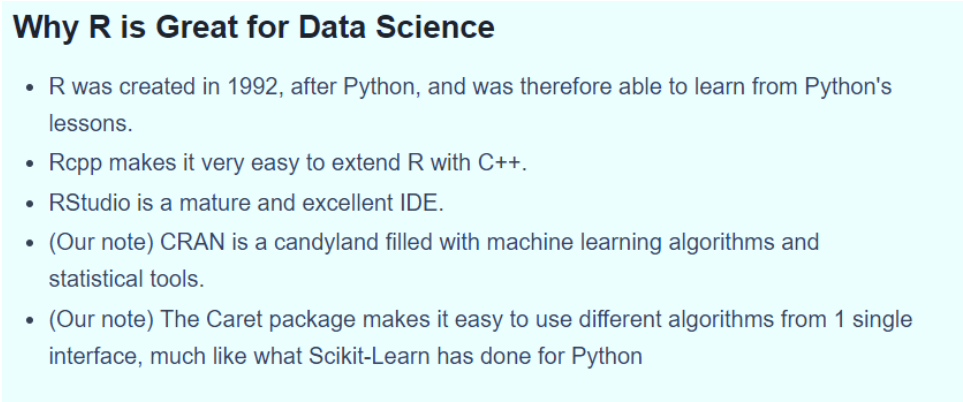

Figure 7: R Studio for Data Science.

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

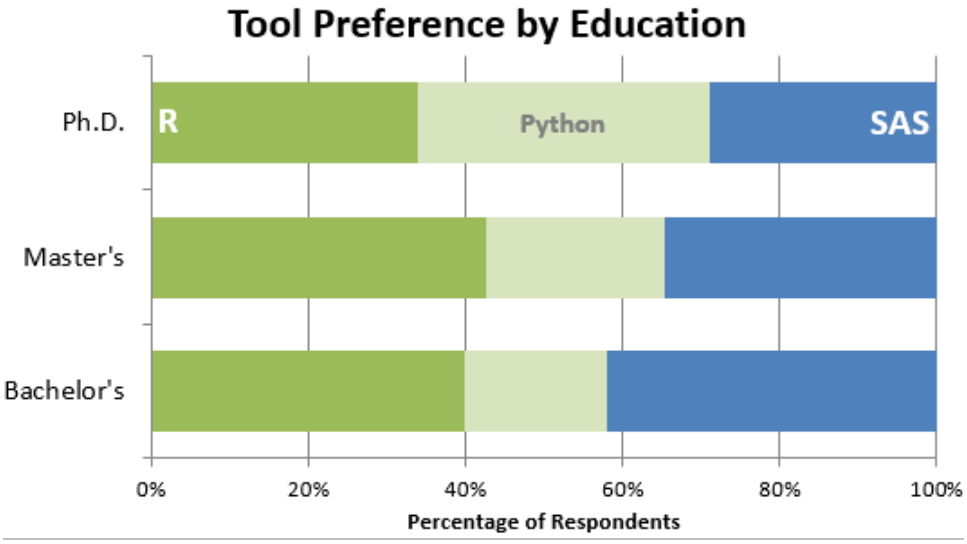

Figure 8: Preference by Education.
(http://www.burtchworks.com/2017/06/19/2017-sas-r-python-flash-survey-results/)

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.

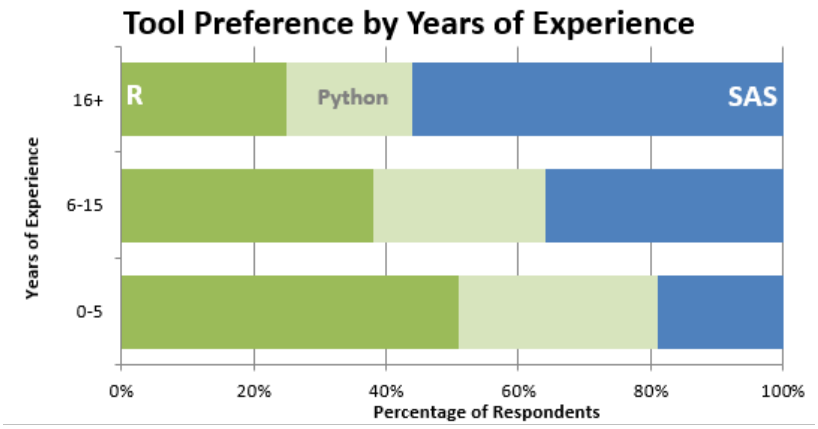

Figure 9: Depicts tool preference by the years of experience.
(http://www.burtchworks.com/2017/06/19/2017-sas-r-python-flash-survey-results/)

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.

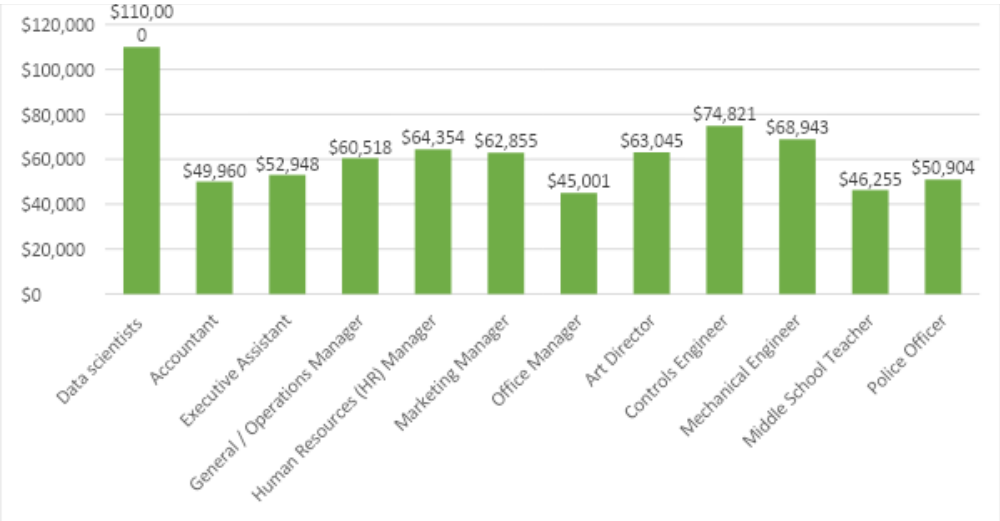

Figure 10: Comparison of salaries of different disciplines against data scientists (Data analyzers)

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.

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.

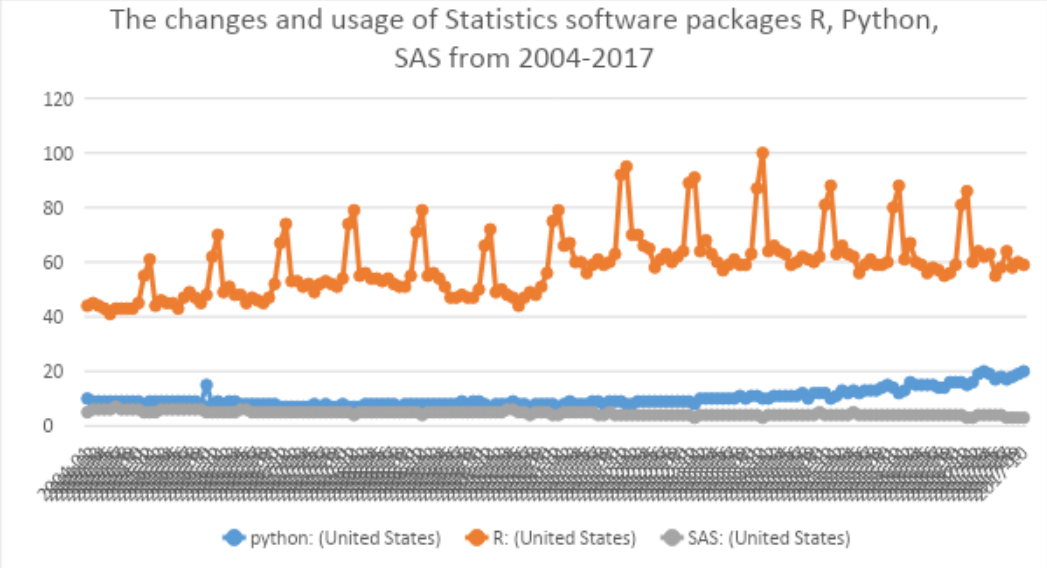

Figure 11: The use of software packages and how they have changed over the years.
(https://trends.google.com/trends/explore?date=all&geo=US&q=python,R,SAS)

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.

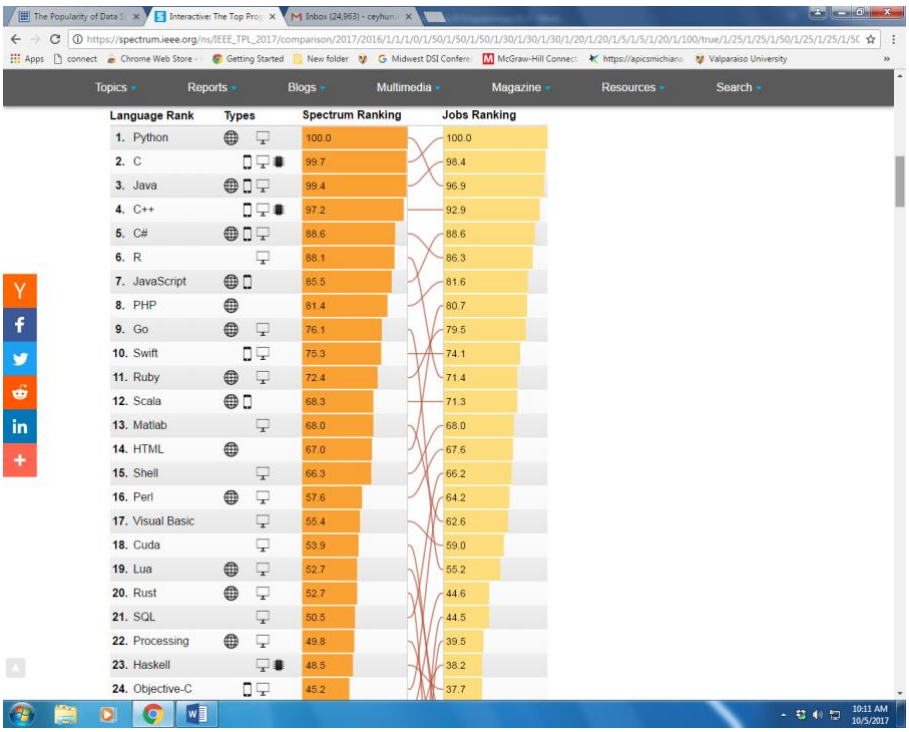

Figure 12: Spectrum and Jobs Ranking of software languages in 2017
(https://spectrum.ieee.org/ns/IEEE_TPL_2017/comparison/2017/2016/1/1/1/1/1/50/1/50/1/50/1/30/1/30/1/30/1/20/1/20/1/5/1/5/1/20/1/100/true/1/25/1/25/1/50/1/25/1/25/1/50/1/25/1/25/1/100/1/100/1/25/1/40/)

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.
Databases such as Oracle should be investigated and included in this study.

References

Ozgur, C., Alam, P., and Booth, D., Software Languages for Analytics in Research. 50th Annual Conference of the Decision Sciences Institute.

SAS Institute Inc., System Requirements for SAS® 9.3 Foundation for AIX®, Cary, NC: SAS Institute Inc.,2012.

2001-2020.Foundation LegalPrivacy Policy Powered By Heroku https://www.python.org/psf/sponsorship/sponsors/

David E. Booth, Ceyhun Ozgur. The use of Predictive Modeling in the Evaluation of Technical Acquisition Performance using Survival Analysis. Volume 17, Number 3, July 2019.

http://www.burtchworks.com/2017/06/19/2017-sas-r-python-flash-survey-results/2000-2020 R-Tools Technology Inc.

https://www.r-studio.com/Unformat_Help/memory_usage.html

System Requirements for MATLAB R2020B, 1994-2020 THE MATHWORKS, INC.

https//www.mathworks.com/support/requirements/matlab-system-requirements.html

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/

https://www.ibm.com/support/pages/hardware-recommendations-ibm-spss-statistics-software

HTTP://WWW.BURTCHWORKS.COM/2017/06/19/2017-SAS-R-PYTHON-FLASH-SURVEY-RESULTS/

C. Ozgur, T. Colliau, G. Rodgers, Z. Hughes, E. B. Myer-Tyson. MatLab vs. Python vs. R. Journal of data Science (2017, 355-372)

https://trends.google.com/trends/explore?date=all&geo=US&q=python,R,SAS

C. Ozgur, S. Jha, Y. Shen. Comparison and Contrast of Statistics Software Packages including R and Python for Teaching Purposes.

PyCharm (January 2017), Python Developers Survey 2016: Findings, https://www.jetbrains.com/pycharm/python-developers-survey-2016/

https://www.quora.com/Is-Python-faster-than-R

https://trends.google.com

https://www.indeed.com/jobtrends

https://spectrum.ieee.org