Issues in Information Systems
Volume 25, Issue 4, pp. 333-343, 2024

DOI: https://doi.org/10.48009/4_iis_2024_127

About the proliferation of data science programming languages: explanatory study, technological development, and common features

Azad Ali, *University of Fairfax, aali@ufairfax.edu*
Umesh Varma, *University of Fairfax, ucvarm@ufairfax.edu*
Shardul Pandya, *University of Fairfax, spandya@ufairfax.edu*

Abstract

This paper discusses the expansion and proliferation of programming languages that have been developed for data science. The paper gives background information on the technological development that led to the growth and expansion of data science languages. It then explains the common features among these languages that distinguish them from general-purpose programming languages. The paper is intended for professionals experienced in working with general-purpose programming languages and may want to enhance their working knowledge of Data Science Programming Languages. The paper further explains the main features that Data Science programming language offers, which may not be readily available in earlier programming languages .

Keywords: Data Science Programming, Big data programming, statistical programming languages

INTRODUCTION

Discussions about programming languages often lead to the thinking of general-purpose programming languages such as Java, C/C++, or Python, among others. Talking about programming languages that fit certain fields may have originated in the days when there were only a few languages in the market and each language was specialized to be used in one area. COBOL, for instance, was heavily used for business applications, Fortran for scientific applications, and BASIC for teaching and learning beginning programming languages (Gabbrielli & Martini, 2023; Jha et al., 2017).

Over time, the number of programming languages in use has mushroomed, along with the applications that programming languages are specialized in. Included among the special-purpose programming languages are data science and statistical programming languages. These have grown quantitatively and qualitatively, and are taking center stage in rankings, surveys, and different analytical studies (Glenn, 2020). Two questions that could be asked here are, (a) how these programming languages and applications evolve; and (b) what distinguishes these languages from general-purpose programming languages. This study intends to answer these two key questions.

The remaining sections of this paper are divided into the following sections:

- The paper explains the technological factors that led to the evolution and expansion of the use of data science programming languages.

- It explains more about data science as a field of study and as a skill that is sought after in the job market.
- Next, the paper highlights data science programming languages, their main features, and what distinguishes them from general-purpose programming languages.
- Finally, a summary and suggestions for future studies are introduced at the end.

Figure 1 below shows the logical progression of this paper.

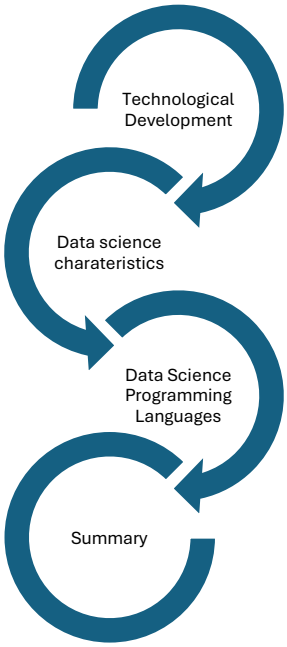

Figure 1: Outline of this study

Data Science Programming Languages Proliferation – Technological Factors

This section discusses factors that contributed to the proliferation of data science programming languages. We start by describing the tremendous increase in data exchanged over the Internet. This increase has been described as "an Avalanche of data" by some, while "a Tsunami of data" by others. We however prefer to describe it as "the mushrooming of data", because avalanches and tsunamis spawn and travel, leaving turmoil in their wake, while Big Data has grown in multiples, but is here to stay. That said, in this paper, we go on to explain how the storage and memory capacity has increased to respond to the increase in the volume of data.

People often use the term "Easier said than done" when describing events or tasks that are difficult to accomplish, illustrate, or put into practice. This can be said about trying to manage the proliferation of data exchange over the Internet. The difficulty arises from the wide range of data exchanged, the enormity of resources that exchange data on the Internet, and sometimes, the unspecified sources of the data.

Having this knowledge about the volume of data exchange over the Internet, we reviewed Google Images in search of a graphic that reflects the volume of data exchange over the Internet. We found two phrases that attempted to do this: "The Internet Minute" and "Data never sleeps". These two terms are eye-catching and can bring many questions when they are introduced into the discussion to exhibit the volume of data

exchanged over the Internet. We found numerous images that exemplify the enormity of data exchange on the Internet. In these images, digital content is used to describe the volume of data. The images that are created take advantage of every corner and every pixel of the image. Figure 2 is one example of an image from 2022 that exemplifies the data exchange over the Internet in just one minute.

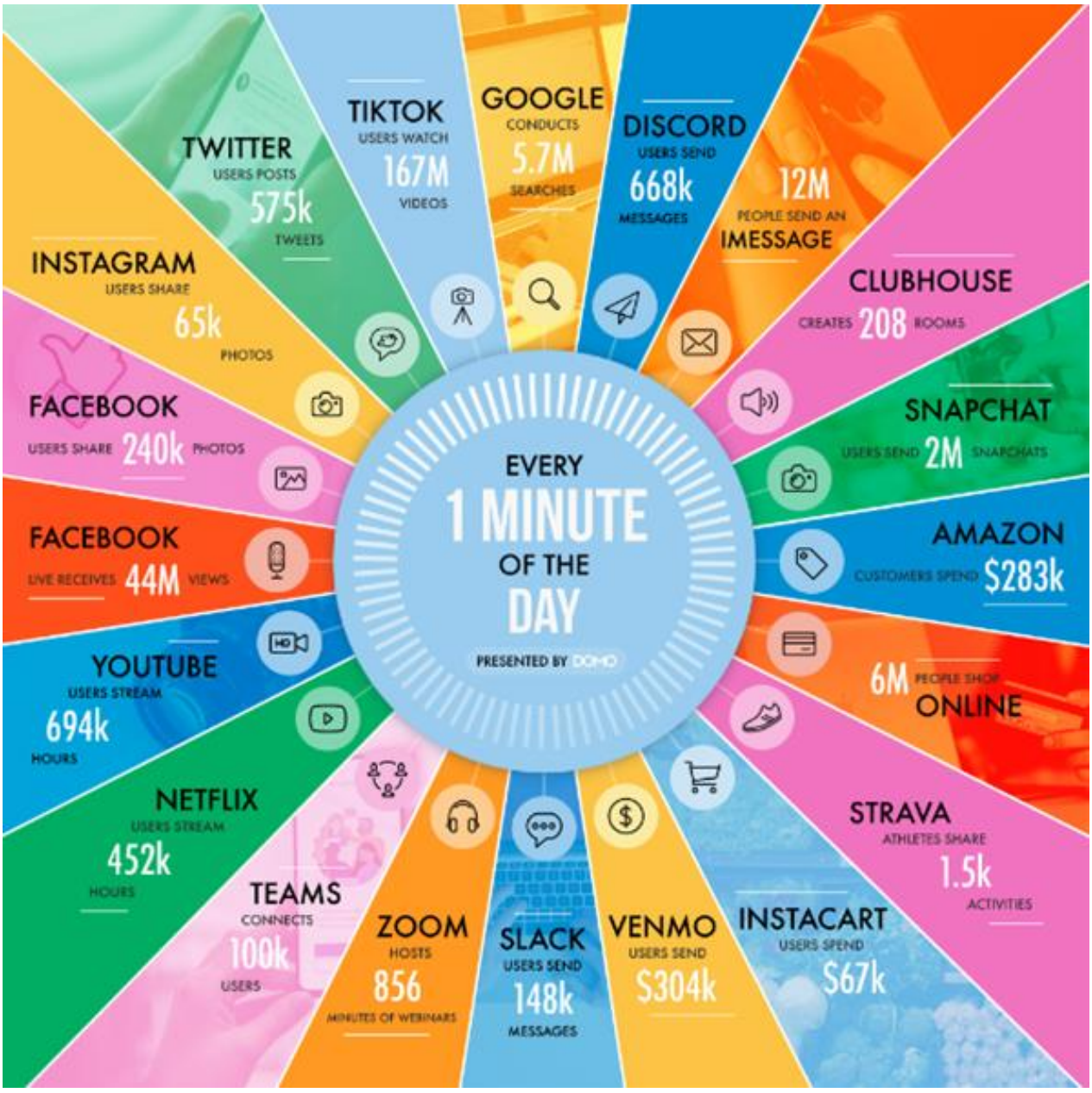

Figure 2: Data Volume Exchanged Every Minute on the Internet (Howlin, 2022)

The figure above shows the enormity of data exchanged globally over the Internet. It does not provide one measure to quantify the volume of data exchanged over the Internet. However, to randomly pick one element from the image above, TikTok users, for example, view 167 million videos per minute. If we multiply this number by the size of a typical video, which runs about nearly 700 MB per video, the result could be more remarkable.

Additionally, we draw attention to these to put data growth in perspective:

- According to one estimate, five exabytes of content were created between the birth of the world and 2003, while in 2013, 5 Exabytes (EB) of content were created each day (Newstex Team, 2014). Today, in 2024, according to the latest estimates, 402.74 million Terabytes (TB) of data are created each day (Duarte, 2024).
    - To put the above in perspective, 5 EB is $5x10^{18}$ bytes, and 402.74 million TB is 4.03×10^{20} bytes. So, the volume of data created each day in 2024 is 3.98×10^{20} bytes more than the total data created until 2013.
- According to Domo's annual report (www.domo.com) a website that specializes in publishing data use over the Internet, they explained that the Internet reached about two-thirds of the world population.
- Statista (https://www.statista.com/) a website that publishes reports about data use over the Internet, noted that data exchanged over the Internet is expected to grow to 181 zettabytes by 2025. That is 1.81×10^{23} bytes.

Suffice it to say that the volume of data exchange over the Internet is staggering. To accommodate this staggering increase in data through electronic data exchange (EDS), we needed storage, memory, and processing capabilities. In the next section, we explore and analyze the proliferation of extensive growth in data storage and memory capacities.

Rapid Rate of Increase in Data Storage/Memory Capacities

The enormous data exchange over the Internet necessitated the need for mass-volume data storage and data processing. The idiom "Necessity is the mother of invention" applies. Researchers in academia and industry focused their attention on this problem and the availability of sophisticated storage devices for this enormity of data was available and kept pace with the increased need for these devices.

To explain the rapid increase in volume, people often use the terms "doubled, tripled, quadrupled", and others like "exponential growth" to exhibit the rate of increase. So, they could say for instance "The population of the world doubled over the last 30 years" which gives a picture of the rate of increase, and using these terms gives a clue of the enormity of these increases.

However, the rate of increase in storage and memory capacities in the computer field is beyond these commonly used phrases. Doubling, tripling, or even a tenfold or hundredfold increase cannot accurately paint the picture of data storage capacity increases. The figures mentioned earlier in this paper, TB, EB, and ZB would apply when discussing storage needs as well but are not easy to remember and compare.

We took it to Google Images again to search for a picture that explains the volumes of data storage and the increase in their capacities. Interestingly, we found a picture that somehow depicts this volume increase in data storage. Figure 3 below shows this data storage/memory capacity increase. Counting the number of zeros included after the first comma may show the enormity of the data storage capacity increase. The number of zeros after the comma is increasing steadily and at a faster pace than before. We can also use the term "exceeded expectations" when comparing the rate of increase in data storage capacities compared what was expected from them.

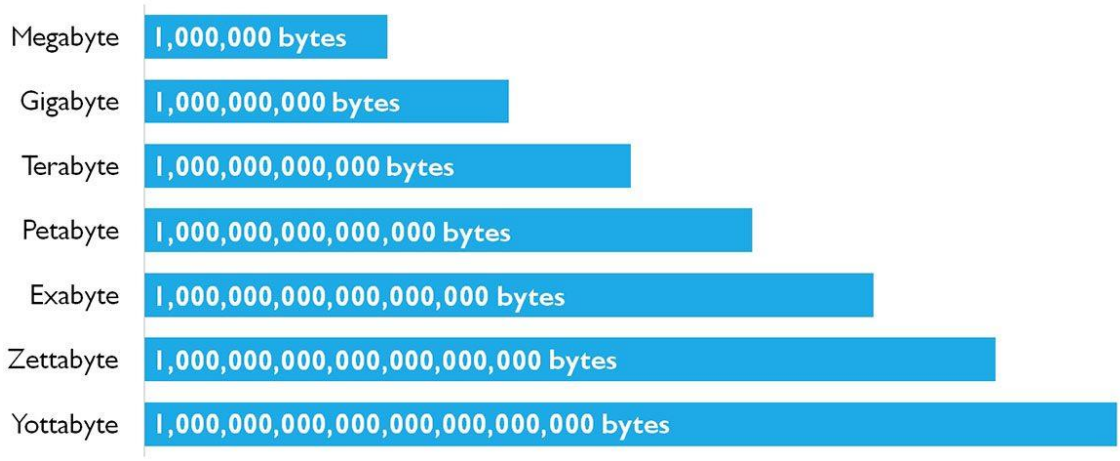

Figure 3: Memory Size Explained (Lynch, 2023)

Data Science Explained

The mass increase in data volume and the need to process and find meaning from the new data measurements were necessitated by a few facts:

- Existing data management technologies are not sufficient to handle this volume of data
- Traditional data processing methods such as batch, real-time, and others do not help with the need for instant processing, or instant reporting.
- Speed of processing and reporting became paramount for finding meaning from the data.
- Existing programming languages were not capable of handling this volume of data and could not take advantage of the knowledge that could be derived from this volume of data.

The changes from the increase in data usage were reflected on two fronts: first, the demand for expertise for data scientists increased. And second, more academic courses and programs were introduced to fill this increasing demand for data scientists.

Data Science jobs

The appearance of data scientists in job ads and the increased demand for people with data science talents started slowly and kept increasing as the volume of data exchange continued to increase. The job pool for data scientists was mixed with statistician, math field, and often computer skills. These developments had a direct impact on the requirements for data scientists and their skills well beyond math/statistics concepts or knowledge of computer programming. This increase in the demand for data scientists is represented by the reports that are published periodically by the Bureau of Labor Statistics, a governmental agency that publishes data about labor market demand and their publications are considered reliable (Bureau of Labor Statistics, 2024).

The Bureau of Labor and Statistics (BLS) publishes annual reports that describe different job titles, the demand for the job, the pay rate, and the expected job growth for ten years at a time. The occupational outlook handbook provided an optimistic prediction of the job demand for data scientists, both in terms of median pay and job outlook for 10 years (2022-2032).

Tablw 1: Data Scientists Job Outlook (Bureau of Labor Statistics, 2024)

Quick Facts: Data Scientists	
2023 Median Pay	$108,020 per year $51.93 per hour
Typical Entry-Level Education	Bachelor's degree
Work Experience in a Related Occupation	None
On-the-job Training	None
Number of Jobs, 2022	168,900
Job Outlook, 2022-32	35% (Much faster than average)
Employment Change, 2022-32	59,400

Other developments contributed to the increasing demand and the recognition of data science as a separate job that needs to be reconciled. Harvard Business suggested that the data analyst's job title is "the sexiest job of the 21st century (Davenport & Patil, 2012). Ten years later, a similar study was conducted to check if data science is still the sexiest job of the 21st century (Davenport & Patil, 2022), and indeed it was! It is interesting to note that in 2024, the topic was still being actively discussed in some quarters, when a similar study was conducted and asked the same questions about data science being the sexiest job (Tuan Zakaria, 2024).

Data Science in Academia

The demand for individuals with the skills necessary to find meaning and intelligence from large volumes of data was unexpected and resulted in a dearth of people with these skills in the market. This prompted academic institutions to begin offering courses and programs that trained students to work as data scientists. The skills required for data scientists' jobs can be found in computer, statistics, math, or similar fields of study. However, there is a clearer distinction of focus between what students are taught when majoring in math, statistics, or computer science – all of which are legitimate foci and necessary for students graduating in those fields – and what data scientists need to know. This distinction necessitated the introduction of new curriculum standards for data scientists (Schwab-McCoy et al., 2021).

The Association of Computing Machinery (ACM) and other organizations periodically publish documents that suggest curriculum content for computing fields of study. The practice began by introducing a standard curriculum for computer science and followed up with introducing other standard curricula for information systems, computer engineering, information technology, information security, and other similar fields of study.

The ACM data science task force developed a document titled "Computing Competencies for Undergraduate Data Science Curricula (ACM Data Science Task Force, 2022) which detailed suggestions for topics to teach, skills to acquire, and many other suggestions to build a program that graduate students in data science. Manifest from the publication of a distinct curriculum in Data Science is the fact that Data

Science is a new and distinct field of study in computing, at par with computing fields suggested by the ACM, such as computer science and information systems. And that data science is different than the fields of math, statistics, and other similar fields of study.

Data Science Programming Languages

The increased volume of data along with increasing storage capacity led to the general recognition of the fact that the value of data has increased. Firms and organizations began to recognize that they could take advantage of the enormous volumes of data available to them. This led to increased investments in storage capacities and simultaneously, the availability of new technologies to store, process, and manage the voluminous data available. The missing key was the technology to process this large volume of data (Pereira et al., 2020).

Data Science programming languages came to the rescue. Although general-purpose programming languages can process some of the data, they do not do it as efficiently. Various vendors wrote code libraries that take the older general-purpose programming languages and generate reports that can be used by data scientists. Different newer statistical programming languages were also developed to fit this gap (Rice University, 2023). Figure 4 shows some of these languages that have been developed along with the general characteristics of each.

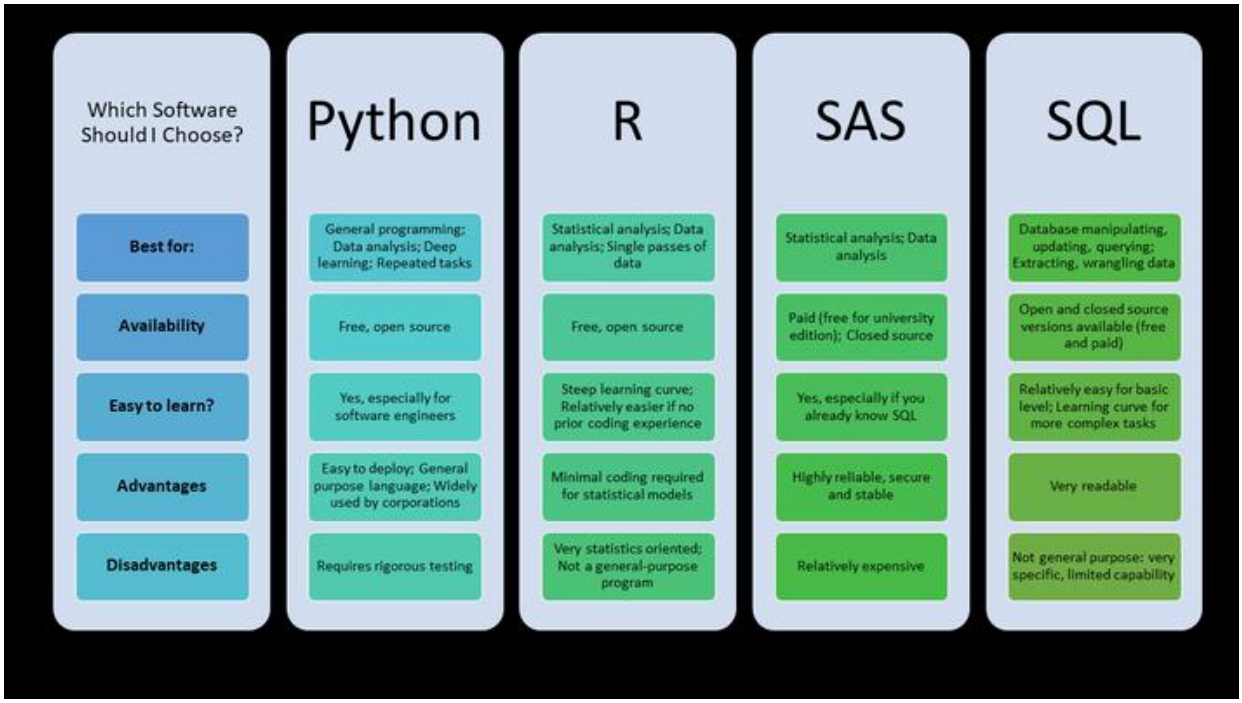

Figure 4 : Example of Data Science Programming Languages (Glenn, 2020)

Common Features of Data Science Programming Languages

Four popular data science programming languages are shown in Figure 4 above. The first programming language, Python is a general-purpose language and is considered a multi-paradigm programming language. Although Python can be found suitable for statistical programming languages (Dhruv et al., 2021), that

specific need can be viewed that it added features to the language and not necessarily built-in features included in the original language.

For this reason, we will focus on the R programming language, which is not considered general purpose and is always ranked high among data science programming languages. Figure 5 below shows an image that explains the different features of the R programming language.

Some features are not very specific to data science. For example, R is open source and easy to use, but this can be said about many other programming languages. However, R has features specific to statistics and data science, such as statistical analysis (including advanced statistical analysis), data manipulation, data visualization, plus outstanding graphical capabilities. The simplicity with which R programming language can complete tasks outweighs the difficult steps that other languages offer them for completing similar tasks (Pavlenko et al., (2022)

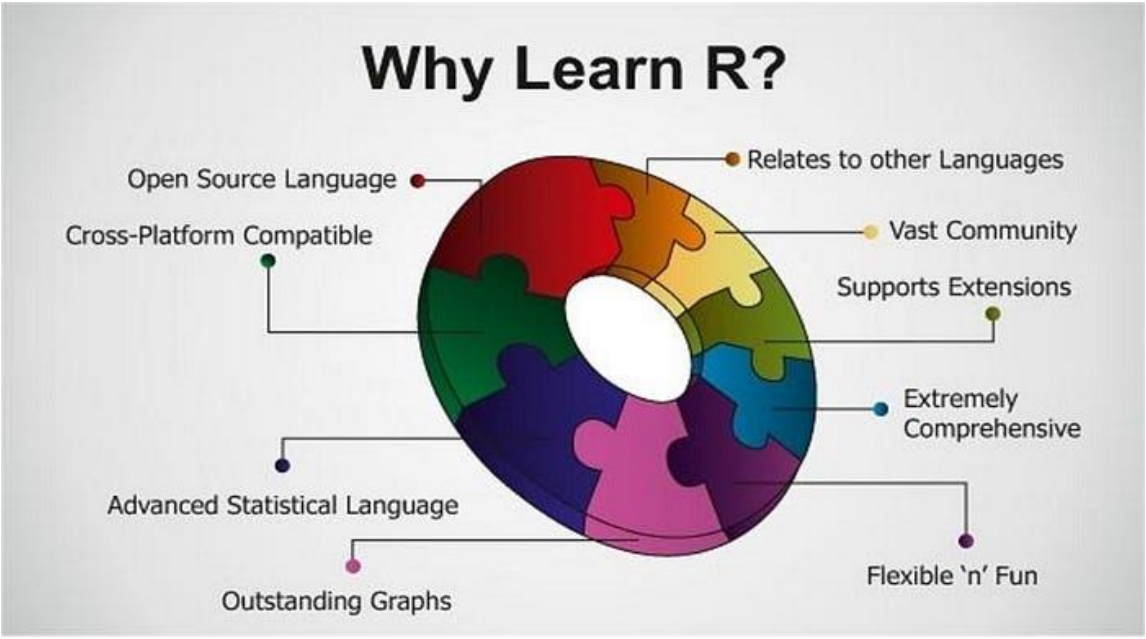

Figure 5: Features of Data Science Programming Languages (Ippolito, 2019)

Researchers distinguish simple statistical analysis, which ranges from simply finding mean, median, and standard deviation to other more complicated ones such as Analysis of Variance (ANOVA), Multiple Analysis of Variance (MNOVA), Generalized Linear Mixed Models (GLMM), and Structured Equation Modeling (SEM). These, and others, can be readily performed in R, by using, as necessitated, appropriate packages and libraries.

In statistics, people often refer to "Descriptive statistics" to suggest providing a summary of data and provide a brief description. This kind of summary, if done in other programming languages could involve using loops, including selection statements, and then finding these measurements such as the mean or standard deviation. However, in data science, this is done with simplicity, and it could include one statement that prints what takes multiple lines of code in other languages (Moon et al., 2023). The same thing can be said about other statistical calculations, that are complicated to complete in general-purpose programming languages, but it is simple and could take only one line of code in data science programming languages.

It was said before that a picture is worth a thousand words, the same thing can be said about a chart is worth a thousand numbers. Charts and graphs provide tremendous advantages to provide a clear and descriptive summary of given data. Indeed, spreadsheet applications such as MS Excel and Google Sheets among others can develop charts with ease but combining that with other programming features and doing it epideictically, needs a programming language with enhanced functionalities to process complex operations with speed, agility, and precision. Achieving such tasks in general-purpose programming languages could take longer steps and multiple modules. While doing them in data science programming languages could be done with little effort and only a few lines of code.

Nordmann et al. (2022) suggested that R programming is full of features that allow easy-to-draw data visualization that shows the direction or trend of data. It cannot be disputed that data visualization provides different advantages for the people who use this platform. However, it is the development of this platform that presents a challenge. Developing a platform that responds visually to users' questions about specific data could take multiple steps and could be complicated. The features in data science programming languages make this task easier. Written reports can be routinely supplemented with different add-on features that show data from various angles and provide the decision-maker with the knowledge/intelligence needed for efficient decision-making. It can be a common routine that written reports can be supplemented with different reports that show data from various angles and provide the decision-maker with the knowledge/intelligence needed for efficient decision-making.

Summary and Suggestions for Future Studies

This paper was about data science programming languages. It started by explaining the technological advances that contributed to the proliferation of these languages and what is termed as "Data Never sleeps". It then discussed data science as a field of study, the job market, and data science programming languages. It elaborated on the features that distinguish these languages from general-purpose programming languages.

While covering the points listed above, we felt that some details were compromised and led to more brief coverage. Based on the knowledge we gained from writing this paper and our diverse expertise in this field, we feel we can offer guidance in developing a program in data science that reflects the latest developments in data science. Since the field of data science is constantly transforming to accommodate the need for complex decision-making, handling live data, creating patterns in real-time, and providing intelligence for knowledge management, the authors recommend the need for further study in the capabilities of data science programming languages.

References

ACM Data Science Task Force. (2022). *.ACM Computing Competencies for Undergraduate Data Science*. https://www.acm.org/binaries/content/assets/education/curricula-recommendations/dstf_ccdsc2021.pdf

Bureau of Labor Statistics. (2024). Occupational Outlook Handbook – Data Scientists. https://www.bls.gov/ooh/math/data-scientists.htm

Davenport, T. H., & Patil, D. (2012). Davenport, T. H., Patil, D. J., & Scientist, D. (2012). The sexiest job of the 21st century. Harvard Business Review, 9. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Davenport, T. H., & Patil, D. J. (2022). Is data scientist still the sexiest job of the 21st century. *Harvard Business Review*, *15*.

Dhruv, A. J., Patel, R., & Doshi, N. (2021). Python: The Most Advanced Programming Language for Computer Science Applications. Cesit 2020, 292–299.

Duarte, F. (2024, June 13). Amount of data created daily (2024). Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day

Gabbrielli, M., & Martini, S. (2023). *Programming languages: principles and paradigms*. Springer Nature.

Glenn, S. (2020, January 27). Best Languages for Data Science and Statistics in One Picture. *Data Science Central*. https://www.datasciencecentral.com/best-languages-for-data-science-and-statistics-in-one-picture/

Howlin, D. (2022). Every Minute – Technology is ticking and producing data 24-7-365. ResearchIP https://research-ip.com/where-was-your-phone-last-night-data-never-sleeps/

Ippolito, P. P. (2019, September 12). Getting Started with R Programming. *Toward Data Science*, https://towardsdatascience.com/getting-started-with-r-programming-2f15e9256c9

Jha, S., Jha, M., O'Brien, L., & Wells, M. (2017, December). Supporting Decision Making with Big Data: Integrating Legacy Systems and Data. In *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)* (pp. 120-128). IEEE.

Lynch, M. (2023, August 5). MEMORY SIZES EXPLAINED: GIGABYTES, TERABYTES, AND PETABYTES IN CONTEXT. Digital and Mobile Technology. *The Tech Advocate*. https://www.thetechedvocate.org/memory-sizes-explained-gigabytes-terabytes-and-petabytes-in-context-2/

Moon, P. F., Israel-Fishelson, R., Tabak, R., & Weintrop, D. (2023, June). The Tools Being Used to Introduce Youth to Data Science. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (pp. 150-159).

Newstex Team (2014, July 12). *The Data Explosion in 2014 Minute by Minute - Infographic.* https://www.newstex.com/blog/the-data-explosion-in-2014-minute-by-minute-infographic

Nordmann, E., McAleer, P., Toivo, W., Paterson, H., & DeBruine, L. M. (2022). Data visualization using R for researchers who do not use R. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221074654.

Pavlenko, L. V., Pavlenko, M. P., Khomenko, V. H., & Mezhuyev, V. I. (2022). Application of R Programming Language in Learning Statistics. In *Proceedings of the 1st Symposium on Advances in Educational Technology* (Vol. 2, pp. 62-72).

Pereira, P., Cunha, J., & Fernandes, J. P. (2020, August). On understanding data scientists. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 1-5). IEEE.

Rice University. (2023, January 20). 12 Best Programming Languages for Data Science and Analytics. Rice University Department of Computer Science https://csweb.rice.edu/academics/graduate-programs/online-mds/blog/programming-languages-for-data-science

Schwab-McCoy, A., Baker, C. M., & Gasper, R. E. (2021). Data science in 2020: Computing, curricula, and challenges for the next 10 years. *Journal of Statistics and Data Science Education*, *29*(sup1), S40-S50.

Tuan Zakaria, T. N. (2024). Is data scientist still the sexiest job for 2024?. *What's What PSPM*.