Python vs Java : A statistical analysis of survey for data science

Rifat Hossain Rafi
ID:a1836604
University of Adelaide
Adelaide, SA, Australia

Abstract

The aim of this research intends to address the conundrum that new learners and rookie programmers experience while deciding on the best programming language for data science. The survey data was used to provide advice based on factors such as age, experience level, language importance in data science, and learning difficulties. Both Python and Java were shown to have advantages in data science; nevertheless, the results show that Python is regarded by industry insiders as being more beginner-friendly. This is due to the fact that it has a clearer syntax and that there are many libraries available specifically designed for data science jobs. The survey's findings demonstrate Python's suitability as the suggested programming language for those just beginning their careers in data science. This experiment offers helpful advice to beginners in selecting their programming language by taking into account elements like ease of learning and the availability of resources.

Keywords: python, java, SQL, survey

1 Introduction

Modern times have seen the emergence of data science as a crucial discipline that combines traditional disciplines like statistics, artificial intelligence, mathematics, and computer technology. It incorporates a number of subdisciplines, including distributed systems, machine learning, and database systems. The main goal of data science is to draw useful conclusions from a wealth of data for the benefit of people, organisations, and society. Businesses and scientific endeavours can make more informed and reliable judgements by utilising data science tools [1]. Due to this, big data science has become increasingly popular. In this field, data science is applied to enormous amounts of data for practical purposes like business optimisation, financial trading, healthcare data analytics, and social network analysis.

Powerful processing systems have been developed to handle the Big Data-related data science activities that are becoming more and more sophisticated. Big Data processing systems have been successful in a variety of domains thanks in large part to the Hadoop framework and its ecosystem [16]. Both Python and Java are frequently utilised for interacting with these data processing platforms. In general, each programming language has its own strengths and weaknesses. We will, however, focus on how they fare in the data science field.

1.1 Background

Some prior knowledge is necessary to comprehend the research question, which is whether Python performs better as a programming language than Java. It is vital to be familiar with the principles, syntax, and applications of programming languages. Understanding Python's and Java's salient traits, as well as those of their respective ecosystems and communities, is helpful.

Python is an interpreted, dynamically-typed language renowned for its ease of use, readability, and robust library support. It is commonly utilised in industries including automation, web development, and data science. On the other hand, Object-oriented programming(OOP) is emphasised in Java, a statically typed, compiled language that is preferred for creating enterprise-level applications. It is crucial to have a fundamental awareness of both languages' advantages and disadvantages. Java is renowned for its scalability, performance, and enterprise integration capabilities, but Python is frequently preferred for its simplicity of use, large libraries, and versatility.

With this background information, one can study the arguments, comparisons, and supporting data for each perspective on the research topic of whether programming language is superior, Python or Java.

1.2 Motivation of Research

In addition to requiring a wide range of sophisticated concepts that can be challenging to understand and apply correctly, programming presents major challenges for beginning students. Programming knowledge beyond its grammatical intricacies is a concept that beginners frequently find difficult to grasp, leading to fragmented comprehension and the inability to combine code fragments into coherent program [17]. Inexperienced programmers frequently write their code in a linear, line-by-line fashion, which hinders their ability to create projects [14]. As a result of the inherent difficulty in learning programming languages, beginners frequently make syntactic or semantic mistakes [8] and hold misunderstandings [13].

Although Java is a popular language for novices, inexperienced students have trouble understanding its difficult

and abstract programming concepts, such as classes, methods, types, and access levels [10]. Python places a strong emphasis on code readability and simplicity in contrast to Java's complex syntax. Python is a good choice as the first programming language because of its succinct syntax, which enables programmers to express concepts in fewer lines of code [6].

The goal of this study is to compare Java and Python in terms of students' comprehension and perception in order to solve the question of which programming language should be taught first to improve novice learners' knowledge while also taking into account both languages' features. This experiment will focus on evaluating not only student's but also professional's comprehension and perception of the two programming languages, aiming to determine which language offers a more favorable learning experience. Evaluation metrics would be based on the popularity, available libraries and the usage of machine learning for these two programming languages.

2 Literature Review

In [5] it indicates that, student's early comprehension of programming is improved by Python's readability and simplicity, but moving to Java is difficult. Clearing up misunderstandings and using targeted instructional techniques helps close the language gap and improve understanding of programming ideas. While Python excels in deep learning packages and user friendliness, Java-based frameworks like Deeplearning4j offer scalability and integration advantages. Depending on the goals of the project, community support, and elements like experimentation (Python) or enterprise scalability (Java), one should choose between the two[9].

In [2] it shows that Java's scalability, performance, and enterprise integration make it the perfect choice for creating big, production-level systems. While Java is prefered for dependable, production-ready big data solutions and compliance with corporate systems, Python is better suited for exploratory research and experimentation.

Aniruddha et al. [12] states taht, Python is a good option for applications in this area due to its vast ecosystem of time series analysis-specific libraries. Python offers tools for data preprocessing, visualisation, forecasting, and anomaly detection using libraries like Pandas, NumPy, and scikit-learn. Java, in contrast, focuses more on enterprise applications and scalability and lacks extensive time series analysis packages. Python has an advantage in time series analysis tasks because to its simplicity, comprehensive documentation, and robust community support.

According to Chieh-An et al.[6] survey findings, the study advises students to take their hobbies and career goals into account while selecting their first programming language. Python's flexibility makes it the perfect tool for experimenting with other subjects, while Java is best for individuals interested in software engineering.

Suarez et al. [15] mentioned about the use of Jupyter Notebooks, which encourage active learning and experimenting with code, formulas, and visualisations, provides a platform for using Python and numerical methods to address structural analysis problems.

Linda et al.[8], analyzed 60 programs written by novice programmers based on a specific range of age in either Java or Python. Later on, analysis of transition from a "simple" language to a more "advanced" one, by following up on couple of HS students, who learned programming in Python before moving on to Java. Linda et al. [3], This process assesses how well a programming language is suited to support both teachers and learners. Next, we explored how programming can be introduced in high school. Since HS students are much younger than college students, this experiment also demonstrates their performance in terms of coding. Edward Raff [11], provided a framework how java can be integrated with machine learning. However, the paper only provides data how to upload the data not about how to build models using java and evaluate the scores. Shuai lu et al.[7], introduced CodeXGLUE framework which benchmarks datasets based on several programming languages. This paper further includes code comparison between python and java. Ben et al.[4], produced a comparison between different programming language regarding their learning difficulties, syntax and running time.

Based on the ideas mentioned above, we gathered a survey kaggle dataset which is 2020 Kaggle Machine Learning Data Science Survey published in kaggle's website. Based on the data provided by the dataset we will try to analyse and provide the insights if python language has the upper hand in comparison with java programming language when it comes to machine learning.

3 Collection of Dataset

3.1 Details of the dataset

A thorough collection of survey results from people actively working in the machine learning and data science fields can be found in the 2020 Kaggle Machine Learning Data Science Survey dataset. The dataset is a useful tool for analysing current trends, practises, and demographics within the business because it includes the thoughts, perspectives, and preferences of over 20,000 respondents.

Figure 1 represents the sample question which were asked to produce dataset from the survey. The dataset includes a wide range of topics, including the tools, frameworks, and programming languages frequently employed by data professionals. It also looks into things like job responsibilities, years of experience, income ranges, and educational background, providing details on the respondents' demographics

Questions	Answers
What is your age ?	Years
What is your gender?	• Man • Woman • Nonbinary • Prefer not to say Prefer to self-describe
In which country do you currently reside?	
What is years the highest level of formal education that you have attained or plan to attain within the nex	• No formal education past high school • Some college/university study without earning a bachelor's degree Bachelor's degree • Master's degree • Doctoral degree • Professional degree • I prefer not to answer
What programming languages do you use on a regular basis? (Select all that apply)	• Python • R • SQL • C • C++ • Java • Javascript • Julia • Swift • Bash • MATLAB • None • Other

Figure 1. Sample of the questions asked during the survey of kaggle dataset

and professional paths.The survey also explores particular data science subfields, such as specialisations, preferred techniques, and industry applications. It gives details on the most common cloud computing platforms, machine learning algorithms, and data visualisation methods used by the respondents.

The dataset enables detailed analysis and exploration of the various factors that contribute to the data science landscape. Researchers and practitioners can use this dataset to gain a deeper understanding of the industry, identify emerging trends, benchmark their own experiences against the wider community, and make data-driven decisions based on the insights derived from this rich dataset.

3.2 Methodology

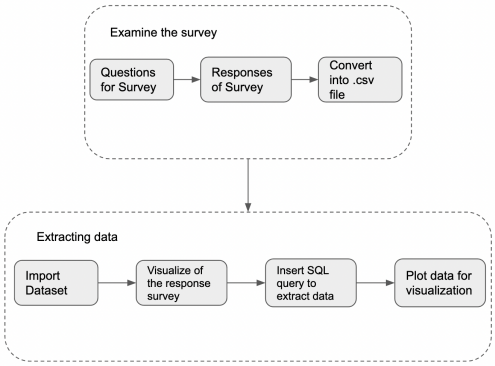

Figure 2. Workflow of the whole experiment about extracting and visualizing the data

Going over the survey's questions and reviewing the responses obtained is the first step. This entails utilising a variety of methodologies and strategies to analyse the information gathered from the survey respondents. In order to ease future analysis, the responses can be transformed into a structured format, such as a.csv file, once they have been collected displayed in Figure 2.

The pandas package from Python is frequently used to import .csv file into a dataframe. Pandas library makes it simple to work with tabular data since it offers effective data structures and data analysis tools. The dataframe makes it simple to explore, manipulate, and visualise the responses gathered from survey.

sql_df = ps.sqldf("Select Q1,Q5,Q6 from df where (Q1 = '18-21' or Q1 = '22-24' or Q1 ='25-29' or Q1='30-34') and Q5 = 'Student' and (Q6 = '< 1 years' or Q6 = '1-2 years')") (1)

Understanding the dataset's structure is crucial after loading the dataset into a dataframe. This entails looking at the columns, the data types, and any empty columns. Structured Query Language(SQL) queries can be used to extract particular data from the dataset. For instance, it can be useful to collect information about participants' years of experience, age, or level of education. Relevant information can be gleaned from the dataset by altering the query in accordance with the desired data.

Once the relevant data has been extracted, it is frequently beneficial to visualise the data for easier comprehension and presentation. The popular Python data visualisation tool matplotlib can be used for this purpose. Plots, charts, and graphs can be made using a variety of visualisation tools offered by Matplotlib to graphically and informatively depict survey data.

The survey data can be efficiently processed, analysed, and visualised by using these procedures, allowing researchers and practitioners to acquire insightful information from the replies gathered.

	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1
0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education ...	Select the title most similar to your current ...	For how many years have you been writing code ...	What programming languages do you use on a reg...
1	1838	35-39	Man	Colombia	Doctoral degree	Student	5-10 years	Python
2	289287	30-34	Man	United States of America	Master's degree	Data Engineer	5-10 years	Python

Figure 3. Display of kaggle survey dataset. Each column represents the response gathered from the participants

3.3 Analysis of the dataset

In order to extract the data for our research we will be focusing on the certain aspect of dataset such as programming

language, years of experience, programming tools and their recommendation in terms of data science/machine learning.

We gather information of the survey from the response of the candidates. Figure 3 shows the first few lines of dataset uploaded into the pandas dataframe.

From the dataset we will split the analysis based on the age group of school and university, their preferred language and the tools they have been using. We will focus on three groups who are in the bachelor's degree, doing college degree and no formal education past high school.

Table 1. Usage of SQl query to gather data based on age, education and experience

Criteria	Data Range
Age	18 - 34
Education	Student
Experience	0 < 2 years

We use SQL query from equation 1 which is straight forward. It gathers data based on specific age range of 18-34 who are currently student and they have few years of experience in terms of programming which is found in Table 1. As it can be seen from the figure majority the students fall into 18-20 age group and they have either less than 1 years of experience or 1-2 years of programming experience. Figure 4 portrays the data of education and experience of programming based on age.

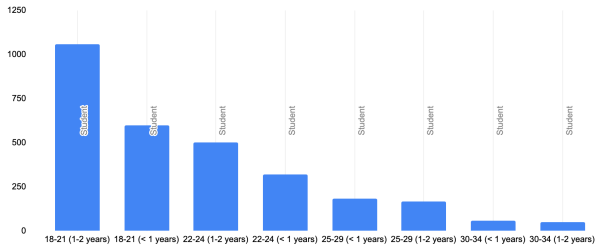

Figure 4. Query based on the age range and education status of the participants

On a regular basis these group of people are using different types of languages such as python, c, c++, java and so on. Since, our experiment focuses on two major programming language for data science projects. Figure **??** shows the statistical analysis of user's prefer java or python. The preferences of python programming language according to the survey is much more higher in comparison with java. This might vary due to several reason. Mainly due to to the availability of python libraries for data science.

Figure 5 represents the easiness of learning programming language between java and python. More than 1000 participants stated that python was easier to learn in comparison

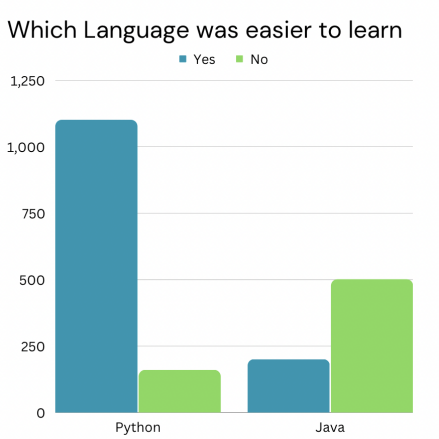

Figure 5. computing platform use most for data science projects between java and python programming language

with java. This is mainly due to easy syntax and less dependency of the libraries. Whereas in terms of java less than 300 participants mentioned that it was easy to learn. However, more than double the participant stated java was not easy to learn.

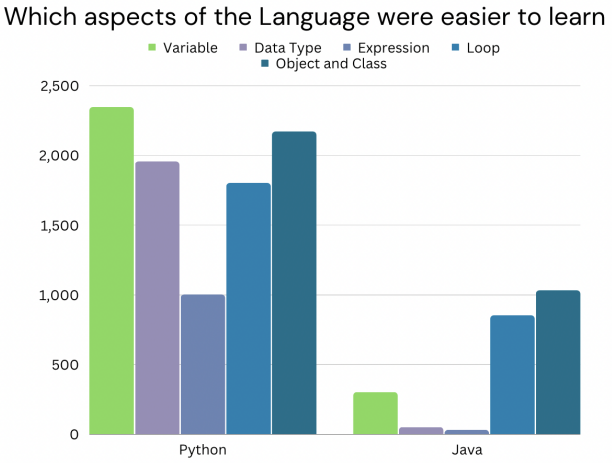

Figure 6. computing platform use most for data science projects between java and python programming language

Figure 7 displays the parameters which were used for the evaluation of the experiment. We used several paramters from each of the language such as variables, loop, class and object, expressions are they difficult to learn or not. Surprisingly, most of the participants with less years of programming language experience stated that, python was easier in comparison with java. On average around 1500 participant mentioned this during the survey.

We also did an analysis on how people with other educational level and many years of experience would recommend

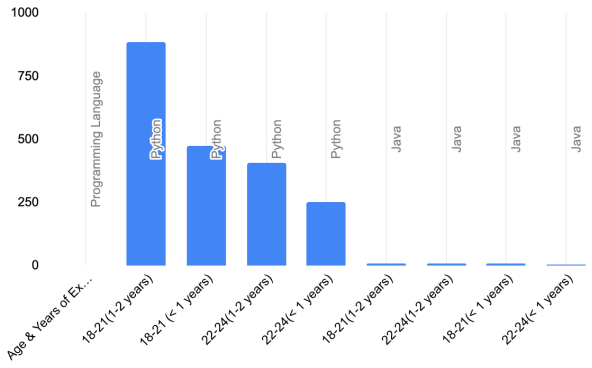

Figure 7. Comparison between java and python based on learning parameters such as variable, loop etc.

programming language for beginners. Most of them recommended python as well. Due to less or no syntax issues and the data are easy to visualize and present to stack holders which can be found in Figure 8.

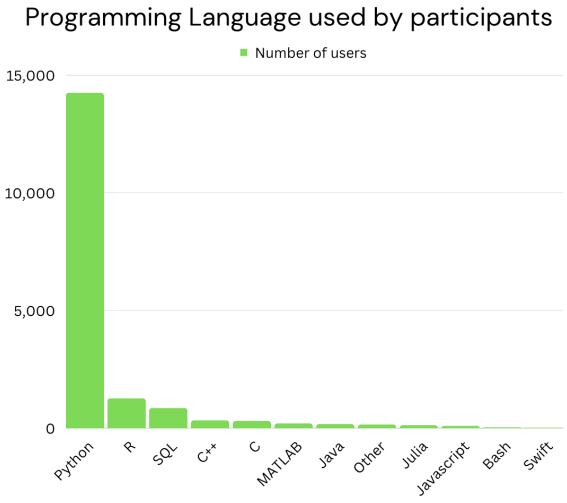

Figure 8. Statistics of most for data science projects based on programming language. This survey is based on responses from all the participants

According to this survey, more than 80% of the participants would recommend python for beginners. Whereas, around 1% participants would recommend java for beginners can be seen in Figure 9. From this statistics we can observe that, in terms of data science, python can be recommend to the beginners for learning purpose.

4 Discussion based on Analysis

From the analysis above we can see that, most of the users who have less than a year of experience tend to shift their focus towards python. The reason might vary depending

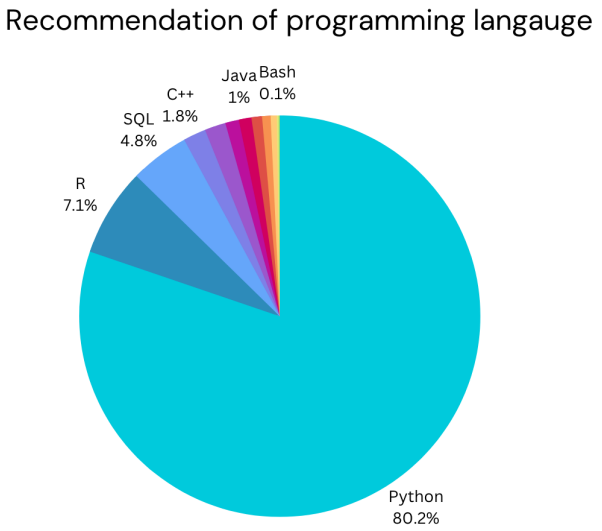

Figure 9. Recommendation percentage of python and java for beginners in terms of data science

on the application they are trying to develop. In terms of data science, python has the upper hand due to easiness of language and availability of vast number of libraries. From the analysis most of the professionals would also recommend python to the beginners. The main reason behind is that, coding is python is easier in comparison with java due to its simpler syntex, variable declaration and compiling process. Based on the analysis above, it can be said that, in term of data science python programming language can easily be recommended to the beginners.

According to the findings, Python is often the language of preference for programmers with less than a year of expertise. There are a number of explanations for this observation, which may differ depending on the particular application these people are pursuing. Python has become the most popular programming language in the field of data science, largely because of how simple it is to use and how many libraries and tools are available that are made expressly for data analysis. Python's popularity in data science can be attributed to its ease of use, logical syntax, and vast library ecosystem. These libraries, including NumPy, Pandas, scikit-learn, and TensorFlow, offer strong and effective answers for a variety of data science tasks, from data manipulation and pre-processing to sophisticated machine learning and deep learning techniques. Python's user-friendliness makes it easy for newcomers to pick up the language and get started solving practical data problems in an instant.

The survey study also suggests that a lot of industry experts advise Python as the best language for beginners. This suggestion can be credited to Python's inherent benefits

over languages like Java, especially when used in the context of data science. Python is easier for beginning programmers to understand because of its simplified syntax, dynamic type, and lack of explicit variable declaration. In addition, compared to Java's compilation process, the creating, running, and debugging of Python code is simpler and less time-consuming.

Given the above findings, it can be said that Python is a very good choice as an introductory programming language for those new to data science. It is the best alternative due to its simplicity of use, broad library support, and widespread popularity within the data science community. Python makes it simple for new users to understand fundamental ideas, experiment with data analysis methods, and advance to more complex machine learning and data science topics. A smoother learning curve and easier entry into the fascinating field of data science are made possible by the enormous resources, courses, and community assistance made accessible for Python.

5 Threats to validity

In the analysis presented above, there are several potential threats to the validity of the findings. These threats should be considered when interpreting the results:

1. Sample Bias: Due to its representation of a certain subset of survey respondents, the survey data utilised in the analysis may be biassed due to sample size. The poll respondents' qualities or preferences could not be typical of the total community of data professionals, or they might have different characteristics from the general public. This might restrict how broadly the results can be applied.

2. Self-Selection Bias: People who choose to react might have unique interests or life experiences that set them apart from those who chose not to. This bias may have an impact on the outcomes and may not fairly represent the views and actions of data professionals as a whole.

3. Response Bias: Results could be skewed or erroneous as a result of participants giving answers they believe to be more positive or socially desirable. Furthermore, by influencing how survey participants read and respond to the questions, the wording or format of the survey questions may generate response bias.

4. Contextual Bias: The research results are unique to the data science setting and the comparison of Python and Java as programming languages in this area. The analysis's findings might not hold true or be applicable in other domains or for other uses that are not data science-related.

5. Timeframe Limitation: The 2020 Kaggle Machine Learning Data Science Survey serves as the basis for the study, which is based on data from that particular time period. Since the study was performed, there may have been new developments or trends because the landscape of programming languages and their popularity can change quickly. As a result, the results might not accurately reflect the state of the field at this time.

To avoid overgeneralizing or deriving erroneous conclusions from the analysis, it is essential to understand these challenges to validity. The constraints and context of the survey data and the particular field of data science should be taken into consideration when interpreting the findings, even when they offer insights.

6 Conclusion and Future Work

The analysis reveals that Python, particularly in the field of data science, is preferred over Java as the programming language of choice for people with less than a year of programming expertise. The results underscore Python's benefits for newcomers, including its simplicity and broad library support. Python is popular among data professionals and is suggested for new programmers joining the field due to its simplicity of use and accessibility to strong data science libraries. However, it is crucial to take into account the analysis' limits. The conclusions are based on survey data, which are subject to response bias, sampling bias, and self-selection bias. The insights reached from this particular data science context might not apply to other domains or reflect the state of programming language technology today.

To validate the results and examine the preferences of a larger group of data professionals, further study may require conducting more thorough and representative surveys. In addition, a long-term study might be carried out to monitor how programming language preferences and trends change over time within the data science industry. Additionally, it would be beneficial to look into the specific factors that contribute to the popularity of Python and compare the potential benefits and drawbacks of using both Java and Python in various data science applications. Research might also compare the effectiveness of various programming languages for data science tasks. This would offer a more thorough understanding of the benefits and drawbacks of various languages and aid novice programmers in choosing the best language for their particular needs.

Overall, it appears that Python is a popular choice for newcomers in data science, but more study is needed to improve our comprehension of the factors influencing language selection and to take into account potential biases in future studies.

References

[1] Monya Baker. 2015. Data science: industry allure. *Nature* 520, 7546 (2015), 253–255.

[2] Radwa Elshawi, Sherif Sakr, Domenico Talia, and Paolo Trunfio. 2018. Big data systems meet machine learning challenges: towards big data science as a service. *Big data research* 14 (2018), 1–11.

[3] Linda Grandell, Mia Peltomaki, Ralph-Johan Back, and Tapio Salakoski. 2006. Why complicate things?: introducing programming in high

school using Python. In *ACM International Conference Proceeding Series*, Vol. 165. 71–80.

[4] Ben Johnson and Anjana S Chandran. 2021. 'Comparison between Python, Java and R progrmming language in machine learning. *Int. Res. J. Modernization Eng. Technol. Sci.* 3, 6 (2021), 1–6.

[5] Fionnuala Johnson, Stephen McQuistin, and John O'Donnell. 2020. Analysis of student misconceptions using Python as an introductory programming language. In *Proceedings of the 4th Conference on Computing Education Practice*. 1–4.

[6] Chieh-An Lo, Yu-Tzu Lin, and Cheng-Chih Wu. 2015. Which programming language should students learn first? A comparison of Java and python. In *2015 International Conference on Learning and Teaching in Computing and Engineering*. IEEE, 225–226.

[7] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664* (2021).

[8] Linda Mannila, Mia Peltomäki, and Tapio Salakoski. 2006. What about a simple language? Analyzing the difficulties in learning to program. *Computer science education* 16, 3 (2006), 211–227.

[9] Aniruddha Parvat, Jai Chavan, Siddhesh Kadam, Souradeep Dev, and Vidhi Pathak. 2017. A survey of deep-learning frameworks. In *2017 International Conference on Inventive Systems and Control (ICISC)*. IEEE, 1–7.

[10] Atanas Radenski. 2006. " Python first" a lab-based digital introduction to computer science. *ACM SIGCSE Bulletin* 38, 3 (2006), 197–201.

[11] Edward Raff. 2017. JSAT: Java statistical analysis tool, a library for machine learning. *The Journal of Machine Learning Research* 18, 1 (2017), 792–796.

[12] Julien Siebert, Janek Groß, and Christof Schroth. 2021. A systematic review of python packages for time series analysis. *arXiv preprint arXiv:2104.07406* (2021).

[13] Teemu Sirkiä and Juha Sorva. 2012. Exploring programming misconceptions: an analysis of student mistakes in visual program simulation exercises. In *Proceedings of the 12th Koli Calling International Conference on Computing Education Research*. 19–28.

[14] James C Spohrer and Elliot Soloway. 1986. Novice mistakes: Are the folk wisdoms correct? *Commun. ACM* 29, 7 (1986), 624–632.

[15] Andrés Suárez-García, Elena Arce-Fariña, Maria Alvarez Hernandez, and Milagros Fernández-Gavilanes. 2021. Teaching structural analysis theory with Jupyter Notebooks. *Computer Applications in Engineering Education* 29, 5 (2021), 1257–1266.

[16] Tom White. 2012. *Hadoop: The definitive guide.* " O'Reilly Media, Inc.".

[17] Leon E Winslow. 1996. Programming pedagogy—a psychological overview. *ACM Sigcse Bulletin* 28, 3 (1996), 17–22.