Introducing Community College Students to Data Science through a Scaffolded Summer Experience

Prof. SAGNIK NATH, University of California, Santa Cruz

Sagnik Nath is an Assistant Teaching Professor (tenure track) in the Department of Computer Science and Engineering at the University of California, Santa Cruz. He earned his Ph.D. in Electrical Engineering from Rensselaer Polytechnic Institute in 2020 and his B.E. in Electronics and Communication Engineering from the Indian Institute of Engineering Science and Technology, Shibpur, in 2015. His teaching and research interests span computer architecture, VLSI design, AI literacy, and optimization techniques for small language models, with a particular focus on integrating generative AI into engineering education. He leads projects exploring how embodied AI, data science instruction, and large language models can enhance student engagement and broaden participation in computing.

Dr. Nath is actively involved in the computer science education community, having served multiple times as an Associate Program Chair for the SIGCSE Technical Symposium and as a panel reviewer for the National Science Foundation. He also reviews regularly for leading computer science and engineering education venues, including ITiCSE, CompEd, and ASEE.

Marcella Gomez, University of California, Santa Cruz
Dr. So Yoon Yoon, University of Cincinnati

Dr. So Yoon Yoon is an assistant professor in the Department of Engineering and Computing Education in the College of Engineering and Applied Science at the University of Cincinnati, OH, USA. Dr. Yoon received her Ph.D. in Gifted Education and an M.S.Ed. in Research Methods and Measurement with a specialization in Educational Psychology, both from Purdue University, IN, USA. She also holds an M.S. in Astronomy and Astrophysics and a B.S. in Astronomy and Meteorology from Kyungpook National University, South Korea. Her work centers on elementary, secondary, and postsecondary engineering education research as a psychometrician, data analyst, and program evaluator with research interests in spatial ability, STEAM education, workplace climate, and research synthesis with a particular focus on meta-analysis. She has developed, validated, revised, and copyrighted several instruments beneficial for STEM education research and practice. Dr. Yoon has authored more than 100 peer-reviewed journal articles and conference proceedings and served as a journal reviewer in engineering education, STEM education, and educational psychology. She has also served as a PI, co-PI, advisory board member, or external evaluator on several NSF- and NASA-funded projects.

Introducing Community College Students to Data Science through a Scaffolded Summer Experience

Abstract

Community college students in the USA represent an important but often overlooked talent pool for computing fields. Yet many face barriers such as math anxiety and low self-efficacy, limiting their interest in CS and STEM transfer pathways. To address this, we developed a 10-day online summer program aimed at shifting pre-transfer STEM-focused community college students' perceptions of data science, an interdisciplinary, CS-centric field blending mathematics, programming, and real-world applications. Rather than a mastery-focused bootcamp, the program served as an exposure-based intervention: a scaffolded, supportive experience designed to give students a taste of what to expect in a formal data science education pathway while fostering confidence, curiosity, and conceptual familiarity. The program featured peer mentorship, scaffolded programming exercises, and guided instruction in key topics such as probability, statistics, and machine learning. Most participants had no prior experience with coding or advanced math but reported increased confidence in applying mathematical concepts in a computing context. Peer mentors played a vital role in creating a non-intimidating environment that encouraged engagement. Despite overall success, students still found some topics, particularly probability and regression, challenging. Pre- and post-program surveys analyzed using the Wilcoxon signed-rank test showed statistically significant improvements in self-efficacy and reductions in math anxiety. Future iterations will incorporate more interactive problem-solving and expanded career guidance to better support student transitions to four-year computing programs. This paper offers lessons from the pilot, including design strategies and student outcomes, to inform scalable computing education interventions for underrepresented learners.

1 Introduction

As data science gains prominence across disciplines [1], engineering institutions are integrating data-centric coursework into undergraduate curricula [2]. Ensuring access to this growing field for historically excluded groups is both an educational imperative and a workforce necessity [3]. Its interdisciplinary nature, spanning programming, mathematical modeling, and real-world application, makes it especially attractive to students from varied academic pathways, including Computer Science and Engineering (CSE) [4,5]. Increasing diversity in data science has also been linked to better innovation, performance, and fairness in AI systems [6–8].

Yet many students from underrepresented backgrounds, especially those in community colleges, face persistent entry barriers, including low self-efficacy [9,10], math anxiety [11,12], and limited access to structured learning environments [10,13]. These barriers are especially acute for pre-transfer students navigating non-traditional paths into STEM.

In response, we developed a fully online summer program spanning three weeks, with synchronous lectures held on 10 designated weekdays. The program targeted pre-transfer, underrepresented community college students intending to pursue engineering or computing majors. It was not designed to produce data science experts, but to shift perceptions: to introduce core data science concepts in a structured, scaffolded way that simulates university-style coursework and builds confidence in a low-stakes environment.

The contributions of this work are: (1) a design framework for short-format, perception-shifting interventions in data science; (2) evidence from a pre/post study using Wilcoxon signed-rank tests showing increased self-efficacy and reduced math anxiety; and (3) insights for adapting such interventions across broader equity-focused computing efforts.

2 Background and Related Work

Efforts to broaden computing participation have led to widespread adoption of introductory data science programs in the U.S. [14]. These often emphasize application-driven or project-based learning while minimizing mathematical depth, based on concerns that rigor may discourage historically marginalized students [3,15].

By contrast, our pilot took a different approach: introducing foundational mathematical concepts, such as probability, statistics, and linear algebra [16], within a scaffolded, supportive environment. Our aim was not to avoid math, but to build students' confidence in engaging with it meaningfully. This aligns with findings that mathematical fluency is essential for advanced work in both data science and broader CSE fields [17]. For community college students eyeing STEM transfer, early exposure to math in a computing context can be transformative.

This program was explicitly not designed to deliver comprehensive data science literacy [18], but to seed perception: to provide students with a brief, structured preview of what data science education entails and cultivate curiosity, confidence, and belonging.

To guide this work, we asked:

- How can a short, scaffolded introduction to data science shift students' perceptions of the field and their place within it?
- Can such exposure-based programs measurably improve self-efficacy and reduce math anxiety for underrepresented students?
- What role does structured peer mentorship play in fostering academic and social belonging in online technical education?

These questions informed both the instructional design and the analysis of student outcomes. Theoretical framing drew from self-efficacy theory [10], math anxiety literature [12,17], scaffolding theory [19], and constructivist learning [20].

3 Program Participant Recruitment

The program was delivered fully online over a three-week period, with synchronous lectures held on 10 designated weekdays within that period. Funding from UC Santa Cruz's Engineering Division Office of Access, Belonging, and Inclusion supported a $500 stipend for each participant. This amount reflected compensation for attending the 10 days of live instruction and was benchmarked to cover the cost of two daily meals (lunch and dinner) at $25 per meal. Recruitment began in January 2024 through a flier sent to in-state community college instructors connected to

UC Santa Cruz. Applicants submitted demographic information and answered the prompt, "What draws you to this program?" via a Google Form, due in late May 2024.

Eligibility was limited to legal adults (18+) who were pre-transfer students. A spreadsheet-based system was used to prioritize applicants based on self-identified gender (with priority to women and non-binary individuals), membership in historically excluded groups (e.g., Latinx, Black, first-generation college students), and adult learner status. Responses further received additional weight based on their response to the aforementioned prompt demonstrated intrinsic motivation for computing, per self-determination [21] and expectancy-value [22] theories. Applicants motivated by curiosity or career growth were prioritized over those citing mainly extrinsic incentives. Final selection used timestamp-based rolling admissions. Selected students had one week to complete IRB consent, with non-responses to completing IRB consents resulting in the next candidate being contacted. From 285 applicants, 35 were thereby selected based on these criteria.

Figure 1 shows the intended majors of the selected cohort: most aimed to pursue Computer Science (60.6%) or Computer Engineering (12.1%), consistent with the view that Data Science aligns closely with computing pathways [4]. The cohort was predominantly female-identifying (80%), with ages ranging from 18 to 45 (mean = 25.6, median = 23), and a non-normal distribution ($p < 0.05$), highlighting inclusion of both traditional and non-traditional students.

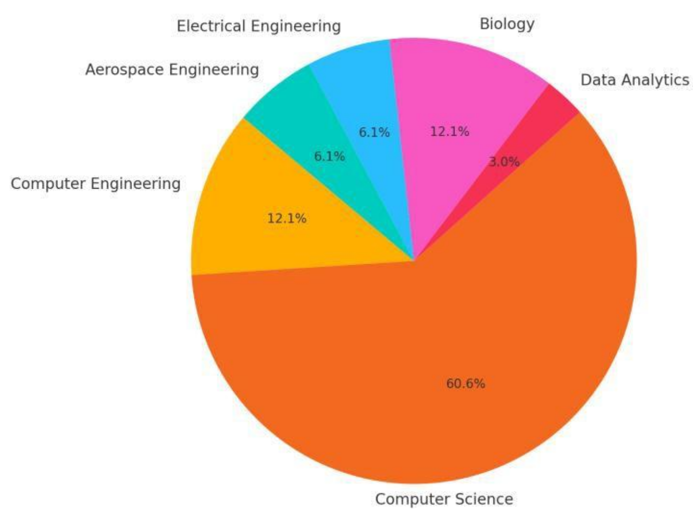

Figure 1: Intended Engineering Majors of Pre-Transfer Community College students participating in the Data Science program

Figure 2 illustrates racial and ethnic diversity: 75.9% identified as Hispanic/Latinx, with smaller proportions identifying as Asian, Black, Multiracial, or other. These demographics align with both the program's equity-focused goals and in-state community college populations [3].

4 Summer Program Design: A Scaffolded, Multi-Modal Experience

This section provides a detailed account of the program structure, instructional elements, and delivery mechanisms employed during the 10-day online summer data science experience. The section further describes the scaffolding techniques that were employed for accessibility, mathematical intuition, and social belonging for underrepresented, pre-transfer community college students.

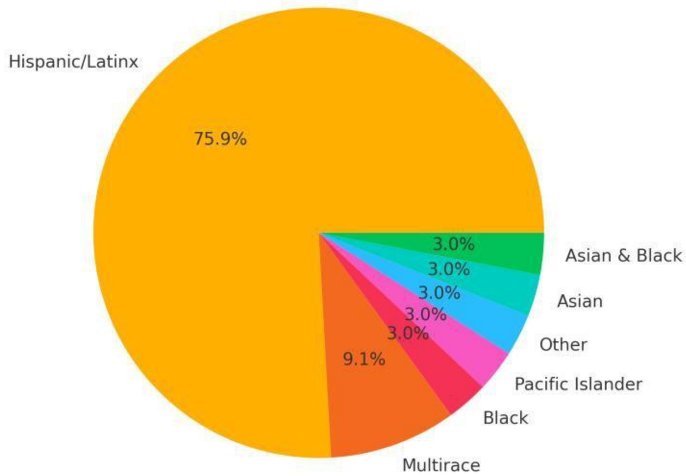

Figure 2: Racial/Ethnic Composition of Participants

4.1 Peer Mentors and Student Support Model

Two undergraduate peer mentors from UC Santa Cruz, both majoring in CSE and having previously completed coursework in Machine Learning and Statistics with an A+ letter grade, were assigned to small student cohorts (17–18 students each). Peer mentors hosted daily breakout discussions during icebreakers, tracked attendance, and held regular office hours. Collectively, they provided five hours of weekly support (in addition to the instructor's three hours), amounting to eight office hours per week. These interactions helped students form connections in a virtual environment and served as key access points for academic support. Mentors also graded homework submissions and provided individualized feedback, reinforcing a non-intimidating, growth-oriented instructional culture

4.2 Lecture Topics and Daily Structure

Before the start of the program, all the students were added to a course Google Drive containing the lecture slides. All the lectures took place in real time online over Zoom. The lectures were also recorded and made available to the students afterwards for download. Each 90 min lecture focused on a core data science concept, progressing from foundational statistics to neural networks. Lectures were structured to begin with an icebreaker (e.g., classification jokes, visual metaphors), followed by a core concept walkthrough, a live demo in Google Colab, and end-of-class discussions. The guest lecture by Amazon data scientist Sabarna Choudhuri on Day 9 added a real-world industry dimension that students cited positively later in surveys.

The following list provides a brief overview of the core topics covered on each day:

Day 1(07/01/24): Orientation and general introductions (announcing name, college, intended transfer major), and answer to an ice breaker question. Then cover course structure, and introduction to data science and machine learning workflows.

Day 2(07/02/24): Introduction to Google Colab environment. A live demonstration of running a simple classifier using the MNIST dataset as an example. Introduction to probability.

Day 3(07/03/24): Introduction to hypothesis testing via real-world testing logic and intuition-building.

Day 4(07/08/24): Z-tests and statistical significance, taught alongside probability distributions.

Day 5(07/09/24): Introduction to Linear regression and statistical modeling using in-state housing prices dataset data in a Colab notebook.

Day 6(07/10/24): Brief overview of Differential Calculus and Linear algebra behind regression and introduction to gradient descent.

Day 7(07/11/24): Bias-variance tradeoff and model generalization using polynomial regression.

Day 8(07/12/24): Logistic regression and binary classification, featuring the diabetes dataset.

Day 9(07/15/24): Career talk from invited Data Scientist from Amazon detailing personal journey into data science (40 mins + 10 min Q&A). Class resumes with precision/recall/F1 metrics for classifier evaluation.

Day 10(07/16/24): Neural networks (from scratch), final synthesis of data science modeling concepts, 15 min presentation from university transfer office representative on applying for transfer process to 4-year university.

4.3 Scaffolded Python Notebooks as Learning Playgrounds

Almost every lecture was paired with a Google Colab notebook, designed to be a hands-on computational workspace. Notebooks emphasized interpretability over syntax, inviting students to modify key code lines to observe the impact of changes.

For example, among the various self-contained python code blocks in the Colab notebook for Day 10, there was one cell that introduced students to visualizing a basic Multi-Layer Perceptron (MLP) neural network. Although the code block itself had various python syntax all students might not have been familiar with (including network visualizations using matplotlib and network Python libraries), their attention was intentionally directed at the following editable line:

layers = [4, 5, 3]

Students were guided to understand that this line defined an MLP with 4 input neurons, one hidden layer with 5 neurons, and an output layer with 3 neurons (Figure 3). Then, during a 15-minute in-class break, students were encouraged to experiment by adjusting these values, for example adding more neurons or multiple hidden layers, to generate real-time neural network visualizations in their own Colab notebooks. This allowed exploration of complex architectures without requiring full coding fluency.

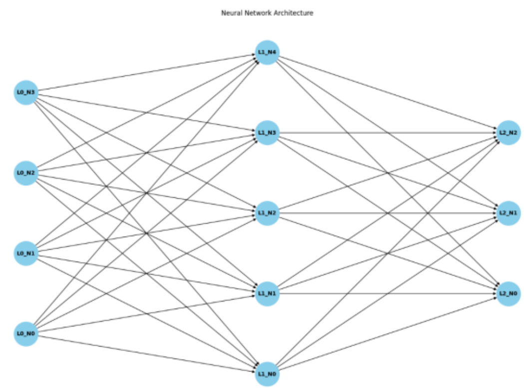

Figure 3: Neural Network architecture visualization tool modifiable to students through altering just one line of code in a Colab notebook introduced on Day 10

One student shared a version of the MLP with over 10 hidden layers with the class, sparking significant engagement. Another remarked enthusiastically on how simply tweaking the layer configuration could dramatically change the complexity of the visualization. This "*control panel*" approach aligns with scaffolding theory and constructivist learning principles, empowering students to learn through guided, hands-on experimentation in a low-risk environment.

4.4 Scaffolded Strategies for Reducing Math Anxiety

To reduce math anxiety and foster early engagement, the program incorporated multiple low-barrier scaffolding strategies. One such approach involved cartoon-style slides (Figure 4) that introduced concepts like hypothesis testing through a humorous, narrative-based progression. Students followed the reactions of an "Armchair Data Scientist" at increasingly surprising results emerging from drug trial data. These panels provided a visual and intuitive entry point into significance testing before transitioning to the formal language of p-values and null hypotheses. By anchoring lessons in relatable storytelling, this constructivist method helped students build conceptual intuition before confronting abstract notation.

Concurrently, math content, particularly linear algebra and differential calculus, was deliberately introduced in minimal, context-driven amount. A preliminary assessment revealed that most students lacked formal exposure to these topics. Rather than omit them entirely, the instruction treated these mathematical ideas as black box tools to aid conceptual learning. For instance, students were shown how matrix multiplications underlie compact linear regression representation, and how simple derivatives help optimize the model parameters. By framing math as a functional tool needed for ***"training"*** models rather than as a mere prerequisite hurdle, students could participate in parameter tuning exercises and later also understand the role of gradient descent without being overwhelmed.

Such layered scaffolds, including visual storytelling, just-in-time math explanations, and selective abstraction, were instrumental in maintaining accessibility while preserving the rigor of the data science concepts taught within the short program time frame.

4.5 Social Belonging Through Scaffolded Icebreakers

Each 90-minute instructional session in the program included a 15-minute peer mentor-led icebreaker designed to foster community and ease students into the day's technical content. One particularly effective session for example was the "Capybara Classifier" icebreaker (Figure 5), which was preceded by a lecture on logistic regression and binary classification. Having just learned how logistic regression naturally lends itself to a binary decision-making framework, students were now geared up to tackle the more abstract idea of extending this to a multiclass setting.

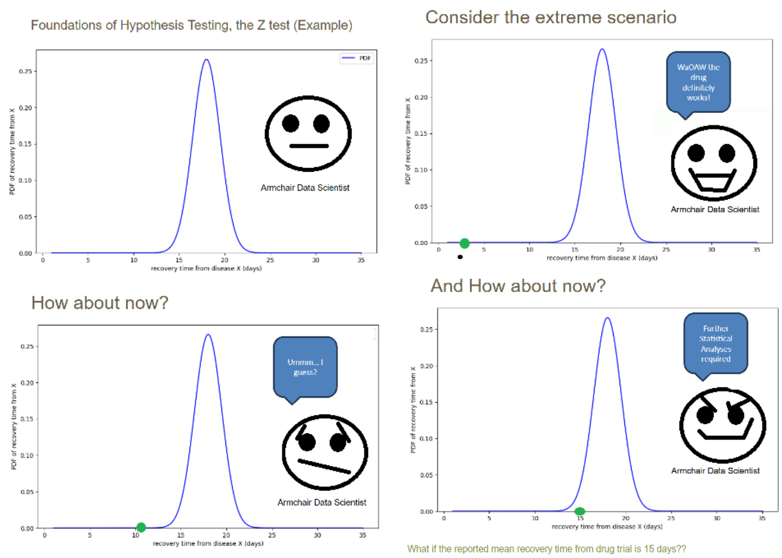

Figure 4: Hypothesis Testing Cartoon Progression captured in 4 sequential slides introduced on Day 3

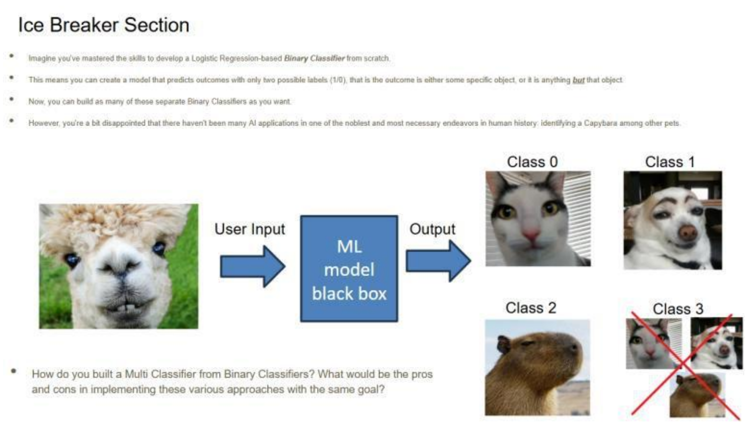

Figure 5: "Capybara Classifier" Icebreaker Slide: using visual humor to scaffold multiclass classification concepts from binary logistic regression

The icebreaker posed a low-stakes challenge of how might one train a machine learning model to reliably identify a capybara among other animals (such as dogs, cats, capybaras or other)? Students were shown an ML "black box" setup in which a user input image would be mapped to one of several classes (Class 0–3), with Class 2 representing the capybara. This visually engaging framing lowered the barrier to participation while encouraging students to build on their existing understanding of binary classifiers. Students were divided into two Zoom breakout rooms, each overseen by a trained peer mentor who had been briefed in advance to steer the conversation in specific directions. Both breakout groups were gently nudged toward deriving the *one-vs-one* (OvO) strategy and the *one-vs-all* (OvA) approach. Rather than directly instructing students on these 2 concepts, mentors offered targeted questions and prompts to help them make the conceptual leap from binary to multiclass classification. The instructor would keep moving between breakout rooms to observe the proper flow of the breakout room narrative. This particular breakout room session saw high levels of engagement from the students who appeared excited to engage with the capybara example and frequently referenced it as a memorable and fun way to understand an otherwise abstract topic. Several participants were able to articulate the advantages and trade-offs of OvO vs. OvA without explicit prior references provided, demonstrating that the intended learning outcomes had been met through this scaffolded, peer-facilitated discussion. They were also encouraged to determine themselves the trade-offs of each method, for example, OvO's computational overhead due to the increased number of classifiers versus OvA's susceptibility to class imbalance. Overall, using this scaffolded peer-led conversations and joyful, relatable framing, we were able to support conceptual development and cultivate a strong sense of belonging within this learning space.

4.6 Scaffolded Homework Assignments

Two scaffolded homework assignments (HW1 and HW2) were issued during the program, reinforcing conceptual and computational ideas taught throughout the lectures and explored in the Colab notebooks. HW1 was designed to build probabilistic thinking through accessible problem contexts, such as a tea-tasting probability exercise inspired by the classic Lady Tasting Tea experiment [23] , and basic code editing tasks involving Python visualizations and memory requirement estimation using the MNIST dataset. This HW emphasized intuition and low-barrier reasoning as students practiced interpreting probability scenarios and modifying simple code cells.

HW2 transitioned to applied modeling. Students ran and adapted starter code on Colab notebooks to perform Z-tests on sample data, trained and visualized simple linear regression models using real-world advertising data, and compared gradient descent and OLS implementations on toy datasets. These exercises drew directly from lecture content and Colab notebooks, with solution strategies scaffolded in example code cells that students could adapt. For instance, in one task, students adjusted a single line of code from the z test examples notebook file provided to explore one- and two-tailed hypothesis tests on battery life claims.

All homework was submitted via Gradescope and manually graded by the two peer mentors. Students could submit any time before the final deadline, which was one week after the program ended. To promote learning over performance, students were openly encouraged to attend office hours where tutors walked them through problem-solving approaches in detail. This created a supportive, low-stakes environment with high instructional availability, reflected in the fact that all 35 participants completed both assignments and met the minimum 80% performance threshold required for the stipend.

4.7 Attendance and the $500 Participation Award

Students were offered a $500 stipend to compensate for incidental costs during the 10-day live instruction schedule within the overall three-week program. This amount was benchmarked at $50 per day of instruction and cover cost of daily meals. Attendance was taken daily by peer mentors. To qualify for the award, participants had to: (i)attend at least 8 out of 10 lectures, (ii)score at least 80% across HW1 and HW2. At the conclusion of the program, all 35 participants earned the $500 participation award based on the criteria.

5 Program Evaluation and Results

5.1 Method

To better understand students' experiences in the program, we administered IRB approved pre- and post-program surveys using a set of 25 7-point Likert-scale items through Qualtrics. These items targeted three constructs: (a) self-efficacy (17 items across five confidence domains—The Basics of Probability [4 items], Foundational Statistics [3], Linear Algebra behind Machine Learning [3], Machine-Learning Basics [4], Career Kick-start [3]); (b) math anxiety (8 items adapted from Ashcraft [12]); and (c) sense of belonging (nine post-survey only items).

While validated instruments for self-efficacy and math anxiety exist, they are typically intended for longer-term courses or general undergraduate populations. In the current study, items were adapted and developed to align specifically with the short-format nature of the intervention (< 3 weeks) and the data science content covered (e.g., probability, regression, gradient descent). This ensured contextual relevance and meaningful interpretation for the program-specific outcomes. As this paper is framed as an experience descriptive report rather than a full empirical study, the survey instrument is intentionally pragmatic with its purpose to capture meaningful directional change within our short-format context and not to serve as a universally validated scale.

We analyzed the 25 pre/post items with Wilcoxon signed-rank tests, appropriate for ordinal Likert data and small samples, and used Wilcoxon one-sample tests (vs the neutral midpoint 4) for the post-only items. 35 students completed the program, but only $n = 25$ returned both the pre- and post-surveys; these 25 respondents qualified for a $20 gift-card bonus that complemented the $500 participation stipend. We also collected general open-ended feedback from students as well.

The outcomes are discussed in Section 5.2

5.2 Key Outcomes

Table 1 summarizes gains across the five self-efficacy scales (17 items total) and the 8-item math-anxiety scale. As shown in the table, most domains showed positive shifts between the pre- and post-program surveys. For example, agreement with item '*I can understand the linear-algebraic principles behind machine-learning technologies*' increased from $M = 4.80$ to $M = 5.92$, $Z = -2.88$, $p = 0.004$ within the confidence domain of Linear Algebra Behind Machine Learning.

Students' largest confidence gains occurred in the Foundational Statistics ($Z = -3.28$, $p = .001$), Linear Algebra behind ML ($Z = -2.84$, $p = .005$), and Machine-Learning Basics ($Z = -4.38$, $p < .001$) domains, exactly the topics emphasized through live coding, gradient-descent visuals, and repeated practice in Colab. In contrast, the Basics of Probability domain increased only modestly and did not reach significance ($Z = -1.26$, $p = .206$), likely reflecting the smaller number of contact hours devoted to probability relative to other topics. The Career Kick-start domain trended upward (Z = -1.70, p = .089) but suggests that explicit career-readiness content, such as transfer-planning workshops or alumni panels, could be strengthened/implemented in future iterations.

Beyond confidence domains, the intervention significantly reduced math anxiety (Z = -2.70, p = .007). Items tapping fear of "not being a math person" and discomfort with timed assessments both dropped by more than one scale point on average, a change students attributed to the scaffolded lectures and peer-mentor support described in Section 4.

Overall, the program succeeded where scaffolded, hands-on exposure was greatest: statistics, linear-algebra foundations, and ML tools, and meaningfully eased students' affective barriers to math. Addressing probability with additional interactive practice and providing more explicit career-guidance components remain clear opportunities for improvement.

Table 1: Statistical Testing on the Changes in Students' Self-efficacy and Math Anxiety Using Wilcoxon Signed Rank Tests *(n* = 25; the 25 participants who completed both pre- and post-surveys)

Confidence Domain	Pre-Survey		Post-Survey		*Z*	*p*
	M	*SD*	*M*	*SD*		
The Basics of Probability	5.85	0.80	6.00	0.77	-1.26	0.206
Foundational Statistics	4.89	1.16	5.76	0.87	-3.28	0.001
Linear Algebra Behind Machine Learning	5.28	1.23	6.15	0.79	-2.84	0.005
Machine Learning (ML) Basics	5.71	0.87	6.03	0.85	-4.38	< 0.001
Career Kickstart	5.33	0.84	5.64	0.89	-1.70	0.089
Math Anxiety	3.43	1.19	2.95	1.05	-2.70	0.007

To gauge students' end-of-program perceptions of academic and social support, we included nine post-only items. Six items (Q26–Q31) formed an "Academic Support" scale (e.g., "My interactions with engineering faculty and mentors were positive"), while three items (Q32–Q34) captured "Peer-to-Peer Social Experience" (e.g., "I was able to positively interact with other student participants").

Academic Support scores were uniformly high (M = 6.45, SD = 0.93 on a 7-point scale). A Wilcoxon one-sample test against the neutral midpoint (4) done for n=25 confirmed that students rated the program's academic resources and mentoring as strongly positive ($Z = –4.22$, $p < .001$). Social-experience items were slightly lower yet still very favorable (M = 5.65, SD = 1.14) with the median above the neutral point ($Z = –3.89$, $p < .001$). These results reinforce the qualitative feedback that the scaffolded mentoring model and peer-led icebreakers (Section 4) fostered both academic confidence and a sense of community, critical outcomes for an exposure-based intervention aimed at shifting perceptions rather than delivering mastery.

Open-ended survey responses also provided additional insight into student experiences. Students described the program as ***"fast-paced but rewarding,"*** with multiple students highlighting the mentorship structure as essential to their success. One student wrote: ***"At first, I was overwhelmed by the math, but the way it was broken down helped me build confidence. I never thought I'd be interested in machine learning, but now I'm actually considering it."*** Another reflected on the impact of peer mentorship: ***"The mentors made the environment really comfortable. It wasn't just about learning the material, it was about not feeling alone in the process."***

6 Lessons Learned

This program reaffirms that a short, scaffolded intervention can reasonably shift perceptions of data science for under-represented community-college students even without aiming for technical mastery. The statistically significant gains in self-efficacy for statistics, linear-algebra foundations, and ML basics, together with lower math anxiety, show that carefully implemented hands-on exposure is enough to build confidence and curiosity.

Two limitations surfaced. First, probability content covered in fewer contact hours did not improve significantly; students asked for more interactive practice. Second, the "Career Kickstart" domain only trended upward, suggesting that career-readiness needs to be addressed explicitly rather than assumed. Future iterations will therefore re-sequence the three-week schedule into a longer duration program comprising first a Python mini-bootcamp followed by a probability & statistics block, and then the core data-science modules, with added career panels and transfer-planning workshops.

Pacing remains a trade-off. Topics such as over-fitting or back-propagation were included to broaden horizons, not for mastery, but some students found them nonetheless conceptually dense. We will either shorten those segments or embed optional deep-dive labs in future iterations.

Peer mentors and a supportive social structure were found to be critical for student engagement and persistence. Feedback and observations showed that accessible mentorship and community eased challenges and supported growth. Short-format CS interventions should therefore prioritize reflection, belonging, and relational support to strengthen under-represented community-college students' perceptions of data science.

7 Conclusions

This report shows that a short, scaffolded program, designed to raise awareness of what data science involves rather than produce expert practitioners, can boost self-efficacy, reduce math anxiety, and widen participation among under-represented community-college students. Future iterations will lengthen the timeline with preparatory Python and probability modules, expand career-readiness content, and retain the peer mentor-driven framework that proved most effective.

Acknowledgements
The authors extend their gratitude to Sabarna Choudhuri from Amazon for delivering a one-hour career talk on Data Science on Day 9.

References

[1] D. Donoho, "Data Science at the Singularity," Harvard Data Science Review, 6(1), 2024. [Online]. Available: https://doi.org/10.1162/99608f92.b91339ef. [Accessed 10 27 2024].

[2] I. Song and Y. Zhu, "Big Data and Data Science: What should we teach," *Expert Systems,* vol. 33, no. 4, pp. 294-413, 2016.

[3] E. Pearson, V. Latson and P. Morgan, "Promoting Data Science to Underrepresented Communities Through the Use of Active Learning," in *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, 2023.

[4] I. Hassan and J. Liu, "Embedding Data Science into Computer Science Education," in *IEEE International Conference on Electro-Information Technology*, 2019.

[5] T. Donoghue, B. Voytek and S. Ellis, "Teaching Creative and Practical Data Science at Scale," *Journal of Statistics and Data Science Education,* vol. 29, no. 1, pp. 27-39, 2021.

[6] Danyluk, Andrea, Paul Leidig, Scott Buck, Lillian Cassel, A. McGettrick, W. Qian, C. Servin, and H. Wang, "Computing Competencies for Undergraduate Data Science Curricula," Association for Computing Machinery, 2021.

[7] Monroe-White, Thema, Brandeis Marshall, and Hugo Contreras-Palacios, "Waking up to Marginalization: Public Value Failures in Artificial Intelligence and Data Science," in *Proc. 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI)*, 2021.

[8] Tu, X., J. Zou, Su, W. and Zhang, L., "What Should Data Science Education Do With Large Language Models?," Harvard Data Science Review, 6(1) , 2024. [Online]. Available: https://doi.org/10.1162/99608f92.bff007ab. [Accessed 27 10 2024].

[9] J. M. Majer, "Self-efficacy and academic success among ethnically diverse first-generation community college students," *Journal of Diversity in Higher Education,* vol. 2, no. 4, pp. 243--250, 2009.

[10] J. E. Maddux, "Self-efficacy theory: An introduction.," in *Self-efficacy, adaptation, and adjustment: Theory, research, and application*, Boston, MA, Springer US., 1995, pp. 3-33.

[11] R. Cohen and A. M. Kelly, "Mathematics as a factor in community college STEM performance, persistence, and degree attainment," *Journal of Research in Science Teaching,* vol. 57, no. 2, pp. 145-307, 2020.

[12] Mark H. Ashcraft,, "Math anxiety: Personal, educational, and cognitive consequences.," *Current directions in psychological science,* vol. 11, no. 5, pp. 181-185, 2002.

[13] J. Sislin and M. Mattis, Enhancing the community college pathway to engineering careers, National Academies Press, 2006.

[14] Naseri, Md Yunus, Caitlin Snyder, Katherine X. Pérez-Rivera, Sambridhi Bhandari, Habtamu Alemu Workneh, Niroj Aryal, Gautam Biswas et al., "Integrating Data Science Into Undergraduate Science and Engineering Courses: Lessons Learned by Instructors in a Multiuniversity Research–Practice Partnership," *IEEE Transactions on Education,* 2024.

[15] L. E. Burger, "Teaching Data Science Programming Skills to Diverse Student Cohorts," 2022, 2022 IEEE IFEES World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC).

[16] T. Nield, Essential Math for Data Science, O'Reilly Media, Inc., 2022.

[17] S. L. Beilock and E. A. Maloney, "Math Anxiety: A Factor in Math Achievement Not to Be Ignored," *Policy Insights from the Behavioral and Brain Sciences,* vol. 2, no. 1, pp. 4-12, 2015.

[18] Magenheim, Johannes, and Carsten Schulte. "Data science education." In *Encyclopedia of Education and Information Technologies*, pp. 493-514. Springer, Cham, 2020.

[19] L.S. Vygotsky,, Mind in Society: the development of higher mental processes (Cambridge,Harvard University Press)., Cambridge: Harvard University Press, 1978.

[20] S.O. Bada, and, S. Olusegun, "Constructivism learning theory: A paradigm for teaching and learning.," *Journal of Research & Method in Education,* vol. 5, no. 6, pp. 66-70, 2015.

[21] E.L. Deci, and R.M. Ryan,, "Intrinsic motivation and self-determination in human behavior," *Contemporary Sociology. ,* vol. 17, no. 2, p. 253, 1988.

[22] Eccles, Jacquelynne S, "Expectancies, values, and academic behaviors," *Achievement and achievement motives,* pp. 75-146, 1983.

[23] Salsburg, David. , The lady tasting tea: How statistics revolutionized science in the twentieth century., Macmillan, 2002.

[24] S. Nath and S. Yoon, "WIP: Beyond Code: Evaluating ChatGPT, Gemini, Claude, and Meta AI as AI tutors in computer science and engineering education," in IEEE Frontiers in Education Conference
(FIE), 2024