Introducing Women to Data Science: Investigating the Gender Gap in a Learning Initiative on Kaggle

Marlon Twyman
University of Southern California
marlontw@usc.edu

Ann Majchrzak
University of Southern California
majchrza@usc.edu

Abstract

Unlike many STEM fields, data science has emerged with online communities serving as prominent spaces for professional development and learning. This paper explores factors that contribute to gender differences regarding perceptions of satisfaction and difficulty in a learning initiative for data science hosted by the Kaggle community. We investigate multiple factors by surveying 2,707 aspiring data scientists: prior experience and skills, professional role, and communication within a learning community. The primary finding is that, despite an initiative intended explicitly to encourage more newcomers (including women) to engage more intensively in learning data science, women dropped out in larger numbers as the five-assignment initiative progressed. Women professed satisfaction with the initiative despite leaving in larger numbers, suggesting a lower expectation about what they had hoped to gain or accomplish from the initiative. Overall, the findings demonstrate how learning initiatives in technically intensive domains contribute to different outcomes between groups.

Keywords: Gender Gaps, Data Science, Online Communities, Informal Learning, STEM

1. Introduction

Online communities for technical fields, such as data science and software development, have proven to be productive environments for people to develop skills and exchange information on various topics (Faraj et al., 2011; Kraut & Resnick, 2012). In general, technical online communities, such as Kaggle, GitHub, and Stack Overflow, share knowledge with large audiences to achieve various professional goals (Blincoe et al., 2016; Jin et al., 2021; Lee et al., 2022; Tsay et al., 2014). Due to the quantity of information available through the projects, these communities also serve as formal and informal learning environments.

People benefit from learning through online communities because of the rich interactivity available to learners. For example, online communities contain content feeds that provide expertise from people that may be otherwise inaccessible to learners (Leonardi, 2015, 2017). Such interactions benefit data science as an interdisciplinary subject where students have limited exposure in formal learning settings, and curriculum at universities are still emerging and developing (Berman et al., 2018; Fekete et al., 2021; Finzer, 2013). The lack of exposure and unsettled curriculum provides an opportunity for online communities to support learners by providing online learning resources, such as access to cloud computing, datasets, and social networks, which all help learners construct learning environments that facilitate their individual knowledge acquisition processes (Anshari et al., 2016).

While these aspects of online communities are positive contributors to learning experiences, online communities have long demonstrated participation gaps among user populations. For example, skills gaps determine access and participation online and are often associated with demographic characteristics, such as age, which disadvantages a subpopulation who may otherwise benefit from participating in a community (Hargittai et al., 2019; Hargittai & Hinnant, 2008; Hargittai & Shafer, 2006). Gender gaps continue to arise and undermine minorities who attempt to access the resources within online communities. As a result, when gender minorities do not contribute at the same rates as the majority, the broader community does not benefit from their perspectives and activity.

The Kaggle data science online community is the area of focus for the current study. It represents an online community centered around a science, technology, engineering, and mathematics (STEM) field where women are underrepresented. As of 2022, fewer than 20% of data scientists identify as women (Kaggle, 2022). The gender gap in data science is unique from other STEM fields since the gap has emerged with online communities serving as the

URI: https://hdl.handle.net/10125/107247
978-0-9981331-7-1

primary spaces for learning and development. More established STEM fields typically have formal classroom settings in which learners are embedded. Within these traditional learning environments, women can be subjected to negative experiences that reduce their engagement with STEM (Etzkowitz et al., 2000; Margolis, 2002). The gender gap in data science, on the other hand, needs to include considerations from both STEM learning and online communities as part of its explanation.

The current study explores how a gender gap emerges among data science learners. Kaggle hosted a learning initiative ("30 Days of ML [Machine Learning]") in the summer of 2021. At the end of the learning initiative, participants who consistently complete assignments should be more familiar with the functions of the platform, including developing and sharing software code, participating in discussion forums, and competing in data science competitions (Dissanayake et al., 2018, 2019; Jin et al., 2021; Tausczik & Wang, 2017).

Our paper contributes to the literature on information technology, social justice, and marginalized contexts by demonstrating how online communities can replicate participation inequalities that exist in offline STEM learning experiences. Even though online communities benefit learners, they do not necessarily alleviate or improve the experiences of minorities in a given field. Overall, the findings from the study offer explanations of the observed gender gap in data science.

2. Background

2.1. Gender Gaps in Online Communities

Online communities generally provide a broad set of activities for contribution and low barriers of entry. These qualities allow for online communities to establish various types of cultural climates. For example, online communities may create experiences that are democratic or encourage behaviors and practices that reproduce systemic biases by constraining the experiences of their members (Miranda et al., 2016).

Underrepresentation of certain demographic groups (e.g., gender and race) has been an issue since the widespread adoption of the Internet (DiMaggio & Hargittai, 2001; van Deursen & van Dijk, 2014). For example, in open source software development projects, men can represent more than 90% of the active community (Hertel et al., 2003; Lakhani & Wolf, 2005). Women are even underrepresented in popular and highly visible communities, such as Wikipedia. The editor population in Wikipedia has reportedly exhibited a gender gap and biases in the editing process since the community's inception (Hill & Shaw, 2013; Langrock & González-Bailón, 2022; Young et al., 2020).

Other previous research studies suggest that communication style is a factor that hinders women from engaging more in online communities. For example, research on Stack Overflow showed that women are usually reluctant to give or receive negative feedback despite the design of the community fostering a culture towards criticism, which becomes a barrier to engagement in the community (Ford et al., 2016). The study also suggests that some women pretend to be men since they perceive men as being more likely to be treated with respect. Such observed behaviors echo the finding that people who are gender minorities can adopt the communication style of the dominant gender in an online community (Mo et al., 2009).

2.2. Gender Gaps in STEM Learning

The lack of women in STEM has been attributed to multiple categories of issues: for example, gender-based stereotypes, lack of familiarity with relevant skills, fewer role models with similar demographic profiles, negative interpersonal climates, and choices in educational topics (Card & Payne, 2021; Cheryan et al., 2011, 2013; Margolis, 2002; Master et al., 2021). These various types of issues relate to individuals and their relationships to others in the environment. Prior research has suggested that addressing negative interpersonal climates will lead to at least two positive benefits for women in STEM: acknowledgement of women's expertise (Joshi, 2014) and more productive research team collaborations (Nielsen et al., 2017).

Over time, stereotypes about women in the workplace have shifted to consider their fit within "masculine" environments (Diekman & Eagly, 2000). The stereotypes of STEM fields possessing masculine environments influences women's choice to enter a field because they may not identify as people with stereotyped personalities, interests, and consumption choices (Cheryan et al., 2017). As an example, software developers are perceived by others as having specific physical characteristics and fitness levels, being socially isolated, and prioritizing work tasks (Chattopadhyay et al., 2021). In some cases, stereotypes are deterrents to involvement in STEM

education because they suggest women have less ability (Appel & Kronberger, 2012).

Previous research has argued that the masculine environment of STEM fields confers a greater sense of belonging and ability to succeed to men than women (Cheryan et al., 2017). Another reason that female students show a lower interest in STEM is that they have fewer related experiences: at early ages, girls and young women reportedly spend more time playing computer and science-related games, playing with technological toys, and have fewer STEM classes in preparation for college (Card & Payne, 2021; Cherney & London, 2006). These reported trends suggest that interests formed at an early age will accumulate as children develop, which will reinforce stereotypes about girls' interest (or lack thereof) in STEM and contribute to gender disparities in motivation to pursue computer science (Master et al., 2021). Such stereotypes can also influence adults who are already in their careers (Chattopadhyay et al., 2021).

2.3. Research Questions

Based on the related prior research, the current study investigates the following research questions to better understand the gender gap that emerges within a learning initiative for data science. At least two factors are potentially relevant to consider: engagement with the content and the issues facing learners. There are two questions focused on exploring gender differences:

- What differences exist between genders with respect to their productivity in the learning initiative?
- How do perceptions of the learning initiative and satisfaction with the learn initiative differ between genders?

3. Data and Methods

3.1. Research Context

Aspiring to be a data scientist is a multifaceted endeavor concerned with technical skill development, interpersonal communication, and solving practical business challenges (Vaast & Pinsonneault, 2021; Zhang, 2019). These elements are present within the prominent online community for data science, Kaggle (www.kaggle.com), established in 2010 with millions of members. In Kaggle, there are numerous activities that members perform when engaging within the platform: developing software code and sharing computational notebooks with analysis, participating in data science competitions, communicating in discussions, and engaging with content from other members.

The "30 Days of ML" learning initiative is a tutorial series curated by Kaggle employees that encourages learners to work on data science assignments for thirty consecutive days. There were five self-paced assignments provided to learners. While learners were not required to complete any assignment, the tutorials were structured to help learners increase their skills throughout the initiative. The first assignment covered basic Kaggle functionality using data from the Titanic shipwreck (Titanic), the second assignment was a module on the Python coding language (Python), the third assignment was an introduction to machine learning (Intro ML), the fourth assignment was on intermediate machine learning (Intermediate ML), and the fifth assignment was an invitation-only competition (Invite-only Competition) where Kaggle employees selected learners to compete.

3.2. Data Collection

Between 2 September 2021 and 9 September 2021, we collected the study data from a survey administered to members of a Kaggle-hosted Discord channel for the "30 Days of ML" learning initiative, resulting in 2,850 survey respondents. A Kaggle employee emailed a link to over 41,000 participants of the initiative (response rate was approximately 7%). Participants were removed if their age was not disclosed or reported as under 18 years old (101 respondents). Then, to focus on the gender gap between men and women, we exclude people who reported other genders (42 respondents), leaving 2,707 respondents (95%) for analysis. No personally identifiable information was collected from respondents and we do not associate their survey responses with any behavioral data from their accounts.

3.3. Survey Description

The survey was administered at the end of the learning intiative and included 15 questions, covering satisfaction with various aspects of the initiative, assignment difficulty, communication frequency, and demographics. Outcome measures of interest include perceptions of satisfaction, difficulty, and barriers to participation. The measures are briefly described in the following paragraphs.

Questions collected demographics, such as the learner gender, age, and current role. Prior technical

skills and knowledge were assessed with two questions. One question assessed previous coding experience with Python. Another question probed familiarity with machine learning.

We measured satisfaction with three questions relating to overall satisfaction with the initiative, satisfaction with the discussion forum experience (offered through the Discord application), and satisfaction with the experience in an invitation-only competition that served as the final assignment. All three satisfaction questions were measured on a 5-point scale (Extremely satisfied [5], Very satisfied [4], Neither satisfied nor dissatisfied [3], Very dissatisfied [2], and Extremely dissatisfied [1]). Respondents could also share if they did not participate in the competition, and state whether they did not or were unable to join the discussion forum.

Perceived difficulty of the five assignments were measured with four options (Too easy [1], Just right [2], Too difficult [3], and Did not complete [NA]). An open-ended question asked, "*What, if anything, did you find difficult about the assignments?*" Another open-ended question assessed barriers to participation by asking, "*What is your most significant blocker for participating in more competitions?*" Other questions requested general feedback on the learning experience and suggestions for improvement.

4. Results

The results first present details about the community of learners by focusing on demographic distributions and skills. Then, we assess gender differences with respect to perceptions of the learning experience (including difficulty) within the initiative. Lastly, we describe barriers and challenges that learners faced when participating in the initiative.

4..1. Description of Learners

Table 1 shows that women were a minority in the Kaggle learning initiative and underrepresented in every age group. Overall, 81.5% of the sample (2,207) reported "Man" as their gender, leaving 18.5% (500) of the total set of respondents identified as "Women." The largest numbers of men and women were present in the lower age ranges. Men were distributed in the age ranges as follows: 18.5% of all men in the 18-21 group, then 26.4% from 22-29, 24.6% in the 30-39 group, 16.1% in the 40-49 group, 9.5% aged 50-59, and 4.9% were in the 60+ group. The age groups for women followed a similar pattern: 24% of all women in the 18-21 group, 29% in the 22-29 group, 24% in the 30-39 group, 15.4% in the 40-49 group, 5.2% in the 50-59 group, and 2.2% in the 60+ group.

Table 1. Learner demographics

		Gender Distributions (Count and Percent of Overall)			
Age Group	**Group Total**	**Men**	**Percent Men**	**Women**	**Percent Women**
18-21	529	409	18.5	120	24
22-29	727	582	26.4	145	29
30-39	664	543	24.6	121	24
40-49	433	356	16.1	77	15.4
50-59	235	209	9.5	26	5.2
60+	119	108	4.9	11	2.2
Complete Sample	**2,707**	**2,207**	**81.5**	**500**	**18.5**

Table 2. X^2 Two-sample test for comparing equal proportions between genders

Age Group	X^2 Between Genders	95% confidence interval
18-21	7.41**	[-0.10, -0.01]
22-29	1.30	[-0.07, 0.02]
30-39	0.02	[-0.04, 0.05]
40-49	0.11	[-0.03, 0.04]
50-59	8.84**	[0.02, 0.07]
60+	6.41*	[0.01, 0.04]
Note: *** p<0.001, ** p<0.01, * p<0.05		

The combined results from Table 2 and Table 3 suggest that older women are most underrepresented in the sample. Table 2 compares the proportions of gender representation in each age group using a X^2 two-sample test. The test assesses the difference between the proportion of genders in each age group. Specifically, for the 18-21 age group, it tests the significance of the difference between 18.5% of men and 24% of women ($p<0.01$). The only other notable differences between genders appear in the 50-59 ($p<0.01$) and 60+ ($p<0.05$) age groups. Therefore, there are differences between how the genders are distributed in the youngest and oldest age groups; a larger proportion of women than men are among the youngest learners while a larger proportion of older men are learners compared to women. No differences are present between the other age groups.

Table 3. X^2 Two-sample test for equal gender proportions between age groups

Age Group	Percent (%) Women	18-21	22-29	30-39	40-49	50-59
18-21	22.7					
22-29	19.9	1.22				
30-39	18.2	3.36	0.56			
40-49	17.8	3.21	0.69	0.01		
50-59	11.1	13.5***	8.99**	5.99*	4.77*	
60+	9.2	10.1**	7.09**	5.18*	4.46*	0.12
Note: *** p<0.001, ** p<0.01, * p<0.05						

Table 3 shows an additional set of analyses using the X^2 two-sample test to assess the difference of the gender distribution between age groups. The gender distributions in the 50-59 and 60+ age groups are more extreme than in the lower age groups. For example, women are 11.1% of the 50-59 group and 9.2% of the 60+ group. Those percentages are significantly lower than the percent of women present at lower age groups.

Table 4. Current role distribution

		Gender Distributions (Count and Percent of Overall)			
Role	**Group Total**	**Men**	**Percent Men**	**Women**	**Percent Women**
Student	981	771	34.9	210	42
Tech Profession	923	782	35.4	141	28.2
Non-tech Profession	443	361	16.4	82	16.4
Other	284	233	10.6	51	10.2
No answer	76	60	2.7	16	3.2

Table 4 shows that learners mostly identified as either a "Student" or a "Professional in a technology related role (e.g., Data Analyst, Data Scientist, Data Engineer)" (Tech Profession). Most women were students (42%) or in a tech profession (28.2%), and most men were in a tech profession (35.4%) or a student (34.9%). The proportion of women students was higher than the proportion of men ($p<0.01$), and the proportion of men in a tech profession was higher than the proportion of women ($p<0.01$). See Table 5 for comparisons between gender proportions in each role.

Table 5. Two-sample test for comparing job roles between genders

Role	X^2 Between Genders	95% confidence interval
Student	8.50**	[-0.12, -0.02]
Tech Professional	9.17**	[0.03, 0.12]
Non-tech Professional	0.00	[-0.04, 0.04]
Other	0.02	[-0.03, 0.03]
No answer	0.19	[-0.02, 0.01]
Note: *** p<0.001, ** p<0.01, * p<0.05		

4.2. Prior Experience and Skills

Prior experience and skills were assessed along two dimensions: prior coding experience and familiarity with machine learning concepts. Three responses were possible for prior coding experience: "Yes, in Python," "Yes, but not in Python," and "No." Experience with Python is relevant because it is one of the programming languages supported by the Kaggle infrastructure and is one of the most popular languages for data science tasks, such as data manipulation and machine learning. Familiarity with machine learning was measured with a 4-point scale: Not familar, A little familiar, Somewhat familiar, and Very familiar.

Table 6. Contingency table of prior coding experience and familiarity with machine learning

			Prior Coding Experience		
			No	**Yes, but not in Python**	**Yes, in Python**
Machine Learning Familiarity **1: Not,** **2: A little,** **3: Somewhat,** **4: Very**	**Men**	**1**	44	83	193
		2	37	184	705
		3	8	86	578
		4	2	23	264
	Women	**1**	26	30	47
		2	12	48	138
		3	0	21	132
		4	0	2	44
	Total		**129**	**477**	**2,101**

The contingency table in Table 6 shows that most of the learning community had prior coding experience

in Python or at least in another language (95% of sample). Familiarity with machine learning was less common, but notable proportions of men and women were at least "Somewhat familiar with machine learning": 44% of men and 40% of women. Compared to women, a higher proportion of men (13.1% to 9.2%; X^2 =5.35, p<0.05) were "Very familiar," and a smaller proportion of men were "Not familiar" (14.5% to 20.6%; X^2 = 11.0, p<0.001).

4.3. Perceptions of Initiative

Table 7 compares genders based on the number of completed assignments, the difficulty of the five assignments, and three dimensions of satisfaction. On average, men completed a higher number of assignments than women (p<0.01), but the only other notable difference is that women perceived the second assignment focused on the Python programming language more difficult than men (p<0.05). Difficulty was perceived between "Too easy" (1) and "Just right" (2) for the first three assignments, but was between "Just right" and "Too hard" (3) for the last two.

Table 7. Gender differences (Mean, St. Dev., and t-statistic) in perceived difficulty and satisfaction

Variable	Sample	Men	Women	t
Completed Assignments	4.63 (0.85)	4.66 (0.83)	4.51 (0.92)	3.28**
Difficulty				
#1 - Titanic	1.74 (0.50)	1.74 (0.50)	1.76 (0.52)	-0.70
#2 - Python	1.78 (0.49)	1.77 (0.49)	1.83 (0.49)	-2.39*
#3 - Intro ML	1.89 (0.43)	1.88 (0.43)	1.91 (0.41)	-1.35
#4 - Intermediate ML	2.07 (0.45)	2.07 (0.45)	2.08 (0.45)	-0.79
#5 - Invite-only Competition	2.18 (0.49)	2.17 (0.49)	2.21 (0.52)	-1.34
Overall Satisfaction	4.21 (0.71)	4.21 (0.71)	4.19 (0.72)	0.57
Competition Satisfaction	4.01 (0.76)	4.02 (0.76)	3.96 (0.79)	1.43
Discord Forum Satisfaction	3.72 (0.81)	3.72 (0.80)	3.71 (0.84)	0.28
Note: *** p<0.001, ** p<0.01, * p<0.05				

Table 8 compares the assignment completions between genders. The largest percentage differences appear at the later assignments. More men completed all five (p<0.001), while more women completed four (p<0.01) and two (p<0.05) assignments.

Table 8. Two-sample test for comparing number of completed assignments genders

Completed assignments	Men (% men in sample)	Women (% women in sample)	X^2
All five	1766 (80%)	355 (71%)	19.0***
Four	263 (11.9%)	85 (17%)	8.95**
Three	88 (4.0%)	29 (5.8%)	2.82
Two	59 (2.7%)	24 (4.8%)	5.51*
Only One	15 (0.7%)	5 (1%)	0.22
None	16 (0.7%)	2 (0.4%)	0.25
Note: *** p<0.001, ** p<0.01, * p<0.05			

The comparisons between genders with respect to the number of completed assignments provide insights into the experience of learners, but it is helpful to investigate those who did not complete all five assignments. As the difficulty of assignments increased, the number of learners who did not complete the assignment increased as well (Table 9). A higher proportion of women did not complete the last two assignments compared to men. The largest differences between genders with respect to assignment completion appeared for the two most difficult assignments: the Intermediate ML assignment (p<0.05) and the Invite-only Competition (p<0.001).

Table 9. Two-sample test for comparing missed assignments between genders

Missed assignments	Men (%)	Women (%)	X^2
#1 - Titanic	47 (2.1%)	17 (3.4%)	2.33
#2 - Python	49 (2.2%)	11 (2.2%)	0.00
#3 - Intro ML	91 (4.1%)	27 (5.4%)	1.30
#4 - Intermediate ML	180 (8.2%)	59 (11.8%)	6.28*
#5 - Invite-only Competition	389 (17.6%)	131 (26.2%)	18.8***
Note: *** p<0.001, ** p<0.01, * p<0.05			

There were no gender differences regarding the overall satisfaction with the learning initiative. However, there was a difference between those who were "Very satisfied" with the competition: 51.3% of men compared to 45.8% of women ($X^2 = 4.7$; $p<0.05$). The only difference that appeared between genders with respect to satisfaction with the Discord discussion forum was for those who were "Very dissatisfied" with the experience: 1.7% of men compared to 3.4% of women ($X^2 = 5.4$; $p<0.05$).

A substantial proportion of learners did not communicate in the Discord discussion forum (752; 27.8%). However, there was no notable difference in the proportion of men (28.2%) and women (26%) who did not use the forum to communicate. The differences between genders regarding perceived difficulty also followed the same pattern described in Table 7. While some proportion of learners did not participate in the Discord discussion forum, they engaged with other Kagglers through YouTube tutorials and by reviewing archival Kaggle discussion forums on various topics.

Overall, genders were mostly similar with respect to satisfaction and perceived difficulty, but assignment completion was a differentiating factor. Men completed more assignments on average, missed fewer assignments at the highest difficulty levels, and a higher proportion of men completed all five assignments. With the exception of the second assignment, women did not perceive assignments to be more difficult despite completing fewer assignments.

4.4. Barriers and Challenges

To better understand what aspects of the learning experience presented challenges, we reviewed responses to two open-ended questions: *"What, if anything, did you find difficult about the assignments?"* and *"What is your most significant blocker for participating in more competitions?"* These questions provided additional feedback and insights from learners. 873 respondents commented on difficulty while 2,615 commented on the "most significant blocker" using more than one word responses. Similar proportions of men (32.3%) and women (31.8%) commented on difficulty. Commenters completed fewer assignments (mean = 4.55, SD = 0.99) than the total sample (mean = 4.63, SD = 0.86).

There were similarities between men and women in their comments and many comments referenced the difference in complexity between assignments as the initiative progressed. For example, a man aged 30-39 who was a professional in a technology related role commented, "It went from 0 to 100 too quickly. I felt like I skipped a step in between because it got really difficult to follow the Intermediate ML course." A student who was a woman aged 18-21 offered a similar observation, "The fact that there was one a day and then suddenly two or three a day, should've done multiple of the easy ones in one day, but I got stuck on day 12 with 3 exercises, and then it escalated too quickly and I got demoralized." These observations illustrate a potential issue with the structure of the initiative: discomfort with the increasing difficulty.

Discomfort with the difficulty increase was at least partially attributed to the prevalence and encouragement to copy and run pre-written software code in the early assignments. Many participants across age-ranges noted an issue with the approach. Another student who was an 18-21 year-old woman stated, "It felt weird to submit to Titanic competition [Assignment #1] when I didn't even understand how it all worked. Also it was hard to find beginner-level ideas that I could use to improve the model for the invite-only competition." Additionally, a 40-49 year-old man who was a professional in a technology related role noted, "It's one thing to read code and understand it; it's another thing to remember it and use it... where you type in everything; I think it helps build muscle memory to repeat these steps, and not just to copy-paste from StackOverflow every time." Learners were able to progress through the assignments, but they did not feel they were prepared for the competition.

The competition was the most difficult assignment in the initiative, and there was a commonly referenced barrier to participation mentioned by 523 learners (19.3%): time. Respondents acknowledged the time commitment needed to complete assignments, the amount of time needed to run computations, the time allocated for other responsibilities, and procrastination. Learners typically made negative comments about feeling limited by their other responsibilities. For example, a 30-39 year-old man in a technology related role mentioned, "In India, we are expected to work for long hours on weekdays and hence didn't get much time on weekday... wish we had something which was for a longer duration, so that we could just manage with weekends for learning and participating in competitions." A 22-29 year-old woman in a technology related role stated a similar issue, "Personal Issue -> work 10 hours in a day and spent only 2 hours to complete. If I am a student, I will have more time to join it." Others across age ranges faced similar issues regarding the time commitment and expectations placed upon learners of the initiative.

5. Discussion

The current study of the "30 Days of ML" in Kaggle presents insights into the gender gap that manifested within the data science online community. Similar to other STEM fields, women were not as productive and were more susceptible to barriers and challenges associated with the structure of the learning environment (Etzkowitz et al., 2000). More women were enrolled as students and held other roles at comparable proportions to men, many had prior experience with computer programming, but did not have as much familiarity with machine learning. The lack of machine learning knowledge was most relevant for the later assignments where fewer women completed the assignments. Another issue was that many learners did not benefit from communication with the rest of the learning community. The lack of exposure to peers reduces access to resources and knowledge (Anshari et al., 2016; Leonardi, 2017).

A higher percentage of women did not complete all assignments, but many learners regardless of gender expressed facing the same barriers. This pattern suggests that the minority group in the community was more susceptible to the barriers than the majority, which highlights the importance of inclusive learning communities. Instead of "masculine" environments, it is valuable to create climates that encourage multiple communication preferences (Cheryan et al., 2017). Making learning initiatives more collaborative by incentivizing peer support, communication, and assistance could shift the currently perceived climates.

5.1. Practical Implications

For online communities that support learning, it is helpful to structure the learning activities to support as many diverse groups as possible. Aside from Kaggle, any online community could benefit from an analysis of how different gender identities are engaging with their resources and what barriers emerge. Also, given that the competition was the last and least completed assignment, online communities for learning can focus on developing norms that foster social support among community members so that learners can build relationships with others in addition to competing with them (Twyman et al., 2023).

Kaggle has a large community and provides numerous resources for multiple elements related to data science, but learners noted that the difficulty of the initiative increased too drastically before the fourth assignment. Also, Kaggle needs to consider the career stage of their learners. Many only have a finite amount of time budgeted to spend on Kaggle learning activities due to other responsibilities, and Kaggle could attempt to design a more inclusive learning environment that supports learners who may have other obligations.

5.2. Limitations

Limitations of the study include associating responses with activities in the learning initiative, measuring the performance and learning of the participants, and generating deeper insights into learner perceptions. Since the survey was collected in a manner to protect privacy, no usernames were collected and we did not connect survey responses to behavioral data. Connecting behavioral data of learners would have been helpful for gaining a more comprehensive understanding of how learners were engaging throughout the initiative. Also, determining approaches to compare assignments would help explain the perceptions of increasing difficulty in the initiative.

A related issue with measuring the performance and learning stems from the prior limitation as well as our data collection. We did not have access to the learning initiative at its start and could only survey learners at the end of the initiative. As such, we were unable to conduct analysis that leverages comparisons before and after the initiative. Additionally, we did not have access to the assignments to assess performance. Learners self-reported whether they completed assignments, but we were not able to assess completion or quality of the assignments.

5.3. Conclusion

Data science is a STEM field where learning is largely supported by online communities. However, online communities experience gender gaps in participation that affect the activities of a community (Young et al., 2020). In a learning initiative hosted by Kaggle, women were underrepresented and also did not produce the same quantity of work compared to men. By analyzing the responses from thousands of learners, it appears that women were more susceptible to the negative elements of the experience than men were. The increasing difficulty of assignments, design of the learning materials, and time requirement all reflect issues that are present in learning environments for other STEM fields that are not situated in online communities (Margolis, 2002). Men and women encountered similar difficulties and barriers, but being

in a minority group was associated with less productivity. The STEM learning experience in online communities needs to be modified to better support minority groups and improve their outcomes.

6. Acknowledgements

We thank Alexis Cook, Julia Elliott, and Myles O'Neill for helping facilitate data collection for this project. This research was supported by the National Science Foundation (IIS-2104551).

7. References

Anshari, M., Alas, Y., & Guan, L. S. (2016). Developing online learning resources: Big data, social networks, and cloud computing to support pervasive knowledge. *Education and Information Technologies*, *21*(6), 1663–1677. https://doi.org/10.1007/s10639-015-9407-3

Appel, M., & Kronberger, N. (2012). Stereotypes and the Achievement Gap: Stereotype Threat Prior to Test Taking. *Educational Psychology Review*, *24*(4), 609–635. https://doi.org/10.1007/s10648-012-9200-4

Berman, F., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. S. (2018). Realizing the potential of data science. *Communications of the ACM*, *61*(4), 67–72. https://doi.org/10.1145/3188721

Blincoe, K., Sheoran, J., Goggins, S., Petakovic, E., & Damian, D. (2016). Understanding the popular users: Following, affiliation influence and leadership on GitHub. *Information and Software Technology*, *70*, 30–39. https://doi.org/10.1016/j.infsof.2015.10.002

Card, D., & Payne, A. A. (2021). High School Choices and the Gender Gap in STEM. *Economic Inquiry*, *59*(1), 9–28. https://doi.org/10.1111/ecin.12934

Chattopadhyay, S., Ford, D., & Zimmermann, T. (2021). Developers Who Vlog: Dismantling Stereotypes through Community and Identity. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 386:1-386:33. https://doi.org/10.1145/3479530

Cherney, I. D., & London, K. (2006). Gender-linked Differences in the Toys, Television Shows, Computer Games, and Outdoor Activities of 5- to 13-year-old Children. *Sex Roles*, *54*(9), 717–726. https://doi.org/10.1007/s11199-006-9037-8

Cheryan, S., Plaut, V. C., Handron, C., & Hudson, L. (2013). The Stereotypical Computer Scientist: Gendered Media Representations as a Barrier to Inclusion for Women. *Sex Roles*, *69*(1), 58–71. https://doi.org/10.1007/s11199-013-0296-x

Cheryan, S., Siy, J. O., Vichayapai, M., Drury, B. J., & Kim, S. (2011). Do Female and Male Role Models Who Embody STEM Stereotypes Hinder Women's Anticipated Success in STEM? *Social Psychological and Personality Science*, *2*(6), 656–664. https://doi.org/10.1177/1948550611405218

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1–35. https://doi.org/10.1037/bul0000052

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as Dynamic Constructs: Women and Men of the Past, Present, and Future. *Personality and Social Psychology Bulletin*, *26*(10), 1171–1188. https://doi.org/10.1177/0146167200262001

DiMaggio, P., & Hargittai, E. (2001). From the 'Digital Divide'to 'Digital Inequality': Studying Internet Use as Penetration Increases. *Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University*, *4*(1), 4–2.

Dissanayake, I., Mehta, N., Palvia, P., Taras, V., & Amoako-Gyampah, K. (2019). Competition matters! Self-efficacy, effort, and performance in crowdsourcing teams. *Information & Management*, *56*(8), 103158. https://doi.org/10.1016/j.im.2019.04.001

Dissanayake, I., Zhang, J., Yasar, M., & Nerur, S. P. (2018). Strategic effort allocation in online innovation tournaments. *Information & Management*, *55*(3), 396–406. https://doi.org/10.1016/j.im.2017.09.006

Etzkowitz, H., Kemelgor, C., & Uzzi, B. (2000). *Athena Unbound: The Advancement of Women in Science and Technology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511541414

Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge Collaboration in Online Communities. *Organization Science*, *22*(5), 1224–1239. https://doi.org/10.1287/orsc.1100.0614

Fekete, A., Kay, J., & Röhm, U. (2021). A Data-centric Computing Curriculum for a Data Science Major. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 865–871. https://doi.org/10.1145/3408877.3432457

Finzer, W. (2013). The Data Science Education Dilemma. *Technology Innovations in Statistics Education*, *7*(2). https://doi.org/10.5070/T572013891

Ford, D., Smith, J., Guo, P. J., & Parnin, C. (2016). Paradise Unplugged: Identifying Barriers for Female Participation on Stack Overflow. *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 846–857. https://doi.org/10.1145/2950290.2950331

Hargittai, E., & Hinnant, A. (2008). Digital Inequality: Differences in Young Adults' Use of the Internet. *Communication Research*, *35*(5), 602–621. https://doi.org/10.1177/0093650208321782

Hargittai, E., Piper, A. M., & Morris, M. R. (2019). From Internet Access to Internet Skills: Digital Inequality Among Older Adults. *Universal Access in the Information Society*, *18*(4), 881–890.

https://doi.org/10.1007/s10209-018-0617-5
Hargittai, E., & Shafer, S. (2006). Differences in Actual and Perceived Online Skills: The Role of Gender*. *Social Science Quarterly*, *87*(2), 432–448. https://doi.org/10.1111/j.1540-6237.2006.00389.x
Hertel, G., Niedner, S., & Herrmann, S. (2003). Motivation of software developers in Open Source projects: An Internet-based survey of contributors to the Linux kernel. *Research Policy*, *32*(7), 1159–1177. https://doi.org/10.1016/S0048-7333(03)00047-7
Hill, B. M., & Shaw, A. (2013). The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PLOS ONE*, *8*(6), e65782. https://doi.org/10.1371/journal.pone.0065782
Jin, Y., Lee, H. C. B., Ba, S., & Stallaert, J. (2021). Winning by Learning? Effect of Knowledge Sharing in Crowdsourcing Contests. *Information Systems Research*, *32*(3), 836–859. https://doi.org/10.1287/isre.2020.0982
Joshi, A. (2014). By Whom and When Is Women's Expertise Recognized? The Interactive Effects of Gender and Education in Science and Engineering Teams. *Administrative Science Quarterly*, *59*(2), 202–239. https://doi.org/10.1177/0001839214528331
Kaggle. (2022, October 11). *2022 Kaggle Data Science & ML Survey: Data Scientists' backgrounds, preferred technologies, and techniques*. Google Cloud Next '22. https://www.kaggle.com/kaggle-survey-2022
Kraut, R. E., & Resnick, P. (2012). *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press. https://doi.org/10.7551/mitpress/8472.001.0001
Lakhani, K., & Wolf, R. (2005). Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects. In J. Feller, B. FitzGerald, S. Hissam, & K. Lakhani (Eds.), *Perspectives on Free and Open Source Software*. MIT Press.
Langrock, I., & González-Bailón, S. (2022). The Gender Divide in Wikipedia: Quantifying and Assessing the Impact of Two Feminist Interventions. *Journal of Communication*, *72*(3), 297–321. https://doi.org/10.1093/joc/jqac004
Lee, J., Park, H., & Zaggl, M. (2022). When to Signal? Contingencies for Career-Motivated Contributions in Online Collaboration Communities. *Journal of the Association for Information Systems*, *23*(6), 1386–1419. https://doi.org/10.17705/1jais.00765
Leonardi, P. M. (2015). Ambient Awareness and Knowledge Acquisition: Using Social Media to Learn "Who Knows What" and "Who Knows Whom." *MIS Quarterly*, *39*(4), 747–762. https://doi.org/10.25300/MISQ/2015/39.4.1
Leonardi, P. M. (2017). The social media revolution: Sharing and learning in the age of leaky knowledge. *Information and Organization*, *27*(1), 47–59. https://doi.org/10.1016/j.infoandorg.2017.01.004
Margolis, J. (2002). *Unlocking the Clubhouse: Women in Computing*. MIT Press.
Master, A., Meltzoff, A. N., & Cheryan, S. (2021). Gender stereotypes about interests start early and cause gender disparities in computer science and engineering. *Proceedings of the National Academy of Sciences*, *118*(48), e2100030118. https://doi.org/10.1073/pnas.2100030118
Miranda, S. M., Young, A., & Yetgin, E. (2016). Are Social Media Emancipatory or Hegemonic? Societal Effects of Mass Media Digitization in the Case of the Sopa Discourse. *MIS Quarterly*, *40*(2), 303–330.
Mo, P. K. H., Malik, S. H., & Coulson, N. S. (2009). Gender differences in computer-mediated communication: A systematic literature review of online health-related support groups. *Patient Education and Counseling*, *75*(1), 16–24. https://doi.org/10.1016/j.pec.2008.08.029
Nielsen, M. W., Alegria, S., Börjeson, L., Etzkowitz, H., Falk-Krzesinski, H. J., Joshi, A., Leahey, E., Smith-Doerr, L., Woolley, A. W., & Schiebinger, L. (2017). Gender diversity leads to better science. *Proceedings of the National Academy of Sciences*, *114*(8), 1740–1742. https://doi.org/10.1073/pnas.1700616114
Tausczik, Y., & Wang, P. (2017). To Share, or Not to Share?: Community-Level Collaboration in Open Innovation Contests. *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW), 100:1-100:23. https://doi.org/10.1145/3134735
Tsay, J., Dabbish, L., & Herbsleb, J. (2014). Influence of social and technical factors for evaluating contribution in GitHub. *Proceedings of the 36th International Conference on Software Engineering*, 356–366. http://dl.acm.org/citation.cfm?id=2568315
Twyman, M., Murić, G., & Zheng, W. (2023). Positioning in a collaboration network and performance in competitions: a case study of Kaggle. *Journal of Computer-Mediated Communication*, 28(4), zmad024. https://doi.org/10.1093/jcmc/zmad024
Vaast, E., & Pinsonneault, A. (2021). When Digital Technologies Enable and Threaten Occupational Identity: The Delicate Balancing Act of Data Scientists. *MIS Quarterly*, *45*(3), 1087–1112. https://doi.org/10.25300/MISQ/2021/16024
van Deursen, A. J., & van Dijk, J. A. (2014). The digital divide shifts to differences in usage. *New Media & Society*, *16*(3), 507–526. https://doi.org/10.1177/1461444813487959
Young, A. G., Wigdor, A. D., & Kane, G. C. (2020). The Gender Bias Tug-of-War in a Co-creation Community: Core-Periphery Tension on Wikipedia. *Journal of Management Information Systems*, *37*(4), 1047–1072. https://doi.org/10.1080/07421222.2020.1831773
Zhang, V. (2019, August 9). *Stop searching for that data scientist unicorn*. InfoWorld. https://www.infoworld.com/article/3429185/stop-searching-for-that-data-science-unicorn.html