Exploring the Evolution of High School Students' Questions in an Interest-Driven Data Science Curriculum

Xiaoxue Zhou, David Weintrop, Rotem Israel-Fishelson, Peter Moon & Yue Xin
University of Maryland

Abstract: This study investigates the data science inquiry process of high school students from populations historically excluded in computing-related fields. We analyzed 213 student-generated questions from the final project of a newly implemented interest-driven data science curriculum. We used a qualitative analytic approach to identify dominant themes of interest and assess question complexity and scope through four stages of data collection. Findings reveal a shift from descriptive to more complex, evaluative, and exploratory questions. Students asked questions from diverse themes, with music and animals being the most common. These insights highlight the importance of scaffolding, culturally relevant content, and adaptive instructional strategies in data science education to empower students from marginalized backgrounds and foster their engagement and success in the field.

INTRODUCTION

Data science, a rapidly evolving field with vast potential, has garnered increasing interest in recent years, particularly in the context of high school education. This surge in interest aims to equip students with the necessary skills to navigate an increasingly data-driven world (Dasgupta, 2016; Krishnamurthi et al., 2020). It is essential to have the ability to formulate meaningful questions that can be explored and answered through data analysis. As such, data science education aims to provide students with technical skills not only to analyze datasets, investigate phenomena, pursue questions, and draw conclusions based on the data (Weiland & Engledowl, 2022), but also aims to equip students to think critically, make informed decisions, advocate for fair data practices, and contribute to creating a more just and inclusive world (Biehler et al., 2022). Recognizing the importance of interests and inquiry in learning data science, especially for young learners, as emphasized by the Interest-Driven Computing Education Framework (Michaelis & Weintrop, 2022), this research delves into the inquiry process of students from populations historically excluded in computing. Specifically, our study investigated the API Can Code. This interest-driven data science curriculum introduces students to computing concepts by encouraging them to pose questions based on their interests and answer them through programming, data analysis, and visualization using publicly available data sources (Weintrop & Israel-Fishelson, 2024). Central to the design of the curriculum is allowing learners to ask and then try and answer questions on topics of interest. As students' data science skills and knowledge progress, they are invited to revise existing and come up with new questions to pursue, providing a unique opportunity to see who students' questions evolve over time.

The research questions guiding this work are:

1. What are the dominant themes of interest expressed by students from historically excluded populations in their data science questions?
2. How do students' data science questions evolve over the stages of the project in terms of type, scope, and complexity?

To answer these questions, we collected and analyzed 213 student-generated questions from the four stages in the final project of the curriculum and identified dominant themes of interest and thematic patterns through a qualitative analytic approach. We created a codebook for question classification and assessed the questions from three dimensions: type, complexity, and scope. This analytic approach allowed us to track the evolution of student questions across the four data collection time points. Our findings shed light on how interest-driven data science projects can empower students from marginalized backgrounds to explore topics relevant to their lived experiences and communities and inform the design of effective data science curricula for these students.

METHODS

We designed and implemented a three-unit data science curriculum with two 12th-grade classes at a Public Charter High School in a city in the Mid-Atlantic region of the United States. The curriculum introduces students to data science concepts in three phases: the nature of data and how it can be accessed, computational foundations of data science, and data analysis and visualization. In each phase, students can guide some of their data investigations based on personal interests, such as choosing their favorite music artist for data collection. In the final project, students chose a topic, crafted a driving question, identified a data source, and then attempted to answer their question using data science practices, presented visualizations, and other analyses. A total of 23 students consented to participate in the study and completed the final project. Table 1 presents the demographics of these students:

Table 1. Participants Demographics

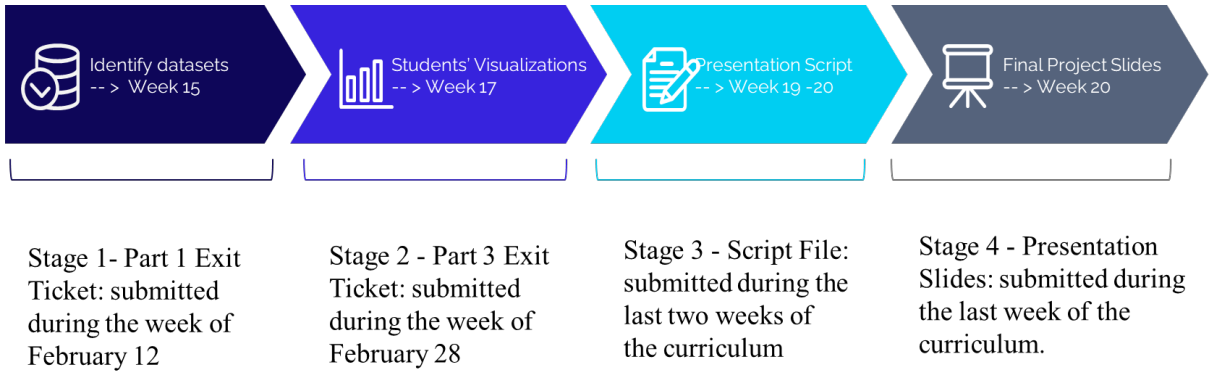
	Total (N)	**Percentage**
Gender		
Female	6	26.09%
Male	17	73.91%
Race/Ethnicity		
Black or African American	20	86.96%
Black or African American, latino	1	4.35%
Hispanic	2	8.7%
Age		
17	10	43.5%
18	13	56.5%

Data Collection Process

The final project asked students to find a dataset on a topic of interest and answer a question using that data. Students accessed API datasets, processed these using programming techniques, and created data visualizations to support their final presentations. Students were prompted to formulate questions at four distinct stages throughout this process, shown in Figure 1:

1. **Identifying Datasets:** Students selected datasets from provided API lists and were asked to "List 4-5 questions you could answer with this data."
2. **Visualization Stage:** Before creating visualizations, students were asked, "What question are you working to answer?"
3. **Sub-Questions for Presentation:** Students created presentation scripts, including primary questions and additional sub-questions.
4. **Final Presentation Slides:** Primary questions were analyzed and answered in the final presentation slides.

Figure 1. Students' Questions Evolution throughout the Four Stages



Questions from the first two stages were collected using an in-class Exit Ticket worksheet. Questions from the third and final stages were collected from the submitted documents for the final presentation.

Data Analysis

We initially classified the collected questions according to their topic. Most questions were aligned with commonly observed themes identified by Israel-Fishelson and colleaguges (2023), such as entertainment, video games, music, and sports.

Subsequently, we developed a codebook to categorize the data from three different perspectives using an open coding technique (Saldaña, 2016). Specifically, our coding manual was informed by typical data science question types (O'Neil & Schutt, 2013; Provost & Fawcett, 2013), and adapted to accommodate the specific characteristics of our collected samples. This resulted in six main question types: Descriptive-Attribute, Descriptive-Comparison, Descriptive-Distribution, Exploratory, Predictive, and Evaluative. The second coding dimension assessed the scope of questions, distinguishing between those narrowly focused on students' interests and those of broader relevance. The final dimension assessed question complexity based on the number of variables involved. Table 2 presents these three dimensions, including definitions and examples from the data. Two researchers independently coded the data (Rivas, 2012). Discrepancies between the two researchers were discussed and resolved), achieving a Cohen's Kappa (Cohen, 1960) of 0.96.

Table 2. Coding Manual

Coding Method	Code	Definition	Examples from Final Projects
Types	**Descriptive-Attribute**	Questions summarizing or quantifying a specific characteristic (attribute) of a dataset	"What is the song length?"
	Descriptive-Comparison	Questions comparing two or more values	"Which dog breed is taller?"
	Descriptive-Distribution	Questions about the frequency or spread of data	"How many movies are comedies?"
	Exploratory	Questions seeking patterns, trends or relationships	"Does the high number of matches played affect goal scores?"
	Predictive	Questions try to predict future outcomes or trends	"How tall can I expect my dog to be?" or "Which team is most likely to win the Super Bowl?"
	Evaluative	Questions assess the value, importance, or effectiveness of something within the dataset.	"What is the best album of 2023 on Spotify?" or "Which album of Jhené Aiko's is most popular?"
Scope	**Broad**	Questions addressing large-scale trends, general characteristics, or aggregate statistics	"How many movies are comedies?" or "What are the most listened to genres on Spotify?"
	Focus	Questions requiring detailed information about specific instances, individuals, or narrow scope	"What awards have Jhené Aiko received?" or "How long can I expect my dog to live?"
Complexity	**Single-variable**	Questions about one specific aspect or dimension of the data.	"How many songs does Jhené Aiko have?" or "What is the artist's name?"
	Multi-variable	Questions requiring analyzing relationships between two or more variables or dimensions within the data	"Does the high number of matches played affect goal scores?"

FINDINGS

We first identified the themes of interest for all 213 questions and created a descriptive table showing the percentage and rank of each category at each stage. Secondly, based on the coding manual, we further categorized the questions into three dimensions: Types, Scope, and Complexity. We then tracked the trajectory of these changes across the four stages as students posed their questions.

RQ1: Dominant themes of interests

As demonstrated in Table 3, the average number of questions per student decreased from 4.65 during the initial identification to 1.42 during the visualization stage. It is then slightly increased to 2.06 and 2.93 during the final presentation stages. When looking at the average percentage of each category across all stages, we found that Music and Animals emerged as the top categories of interest among the students in their data science questions. For instance, the proportion of questions related to Animals increased from 19.63% during the initial identification stage to 33.33% during the visualization stage. Similarly, the interest in Music consistently remained high, peaking at 31.82% during the final presentation slides stage.

Table 3. Dominant Themes

Stages of Students Questions Collected	**Total Number of Participants**	**Total Number of Questions**	**Animals**	**Entertainment**	**Music**	**Sports**	**Video Games**
Identifying Datasets: (week 15)	23	107	19.63%	18.69%	22.43%	21.50%	17.76%
Visualization Stage (week 17) :	19	27	33.33%	14.81%	14.81%	18.52%	18.52%
Sub-Questions for Presentation (week 19-20)	17	35	23.53%	20.59%	26.47%	11.76%	17.65%
Final Presentation Slides (week 20)	15	44	25.00%	13.64%	31.82%	15.91%	13.64%

Note: This table shows that the number of participants decreased, but the average number of questions per student slightly increased by the final presentation as students progressed through the final project. Interest in animals and music remained consistently high across all four stages.

RQ2: Evolution of the data science questions throughout the project

Student-generated questions demonstrate a distinct evolution across the project's four stages (Table 4). Initially, students primarily asked basic, broad, single-variable questions, reflecting a familiarization with the data and their personal interests. For example, "*What is the average age of players in the NBA?*" (Student 6) or "*How many movies are comedies?*" (Student 3). These descriptive questions (44.86%) focused on single variables (93.46%) and indicated initial data exploration. Notably, some students expressed interest in prediction ("*How tall can I expect my dog to be?*" - Student 9), but this type of question disappeared in later stages.

The visualization stage marked a shift towards more focused and comparative inquiries. Descriptive-Comparison questions increased from 19.63% to 44.44%, exemplified by Student 6's shift from "*What species exist in the Star Wars universe?*" to "*What species is the most populated in Star Wars?*" Evaluative questions also grew (7.48% to 14.81%), often using terms like "best" or "popular." Student 4's question about Jhene Aiko's music changed from "*What awards has she received?*" to "*Is her music popular?*"

In the final stages, question diversity expanded, reflecting deeper analysis and a greater focus on specific aspects of the data. Sub-questions for scripts showed increased complexity (14.29%), exploring topics like song duration and explicit lyrics (Student 1) or the relationship between match location and goal scores in sports (Student 4). While descriptive questions remained prevalent, both evaluative (11.36%) and exploratory (11.36%) questions increased compared to the initial stage. Questions like "*Is there a correlation between a team's payroll and their number of wins in a season?*" (Student 12) and "*Which genre of music is most popular among teenagers?*" (Student 5) demonstrate a more critical engagement with the data and an effort to uncover relationships and patterns.

Table 4. Students' Questions Evolution throughout the Four Stages

Stages of Students Questions Collected	Desc. Att.	Desc. Comp.	Desc. Dist.	Eval.	Explor.	Predic.	Broad	Focused	Single-Var.	Multi-Var.
Identifying Datasets (week 15):	44.86%	19.63%	15.89%	7.48%	9.35%	2.80%	65.42%	34.58%	93.46%	6.54%
Visualization Stage (week 17)	22.22%	44.44%	11.11%	14.81%	7.41%	0.00%	74.07%	25.93%	96.30%	3.70%
Sub-Questions for Presentation(week 19-20):	31.43%	34.29%	17.14%	5.71%	11.43%	0.00%	77.14%	22.86%	85.71%	14.29%
Final Presentation Slides (week 20):	34.09%	22.73%	20.45%	11.36%	11.36%	0.00%	59.09%	40.91%	77.27%	22.73%

Note: Desc. Att. = Descriptive-Attribute; Desc. Comp. = Descriptive-Comparison; Desc. Dist. = Descriptive-Distribution, Eval. = Evaluative; Explor. = Exploratory; Predic. = Predictive; Single-Var. = Single Variable; Multi-Var. = Multi Variables

This table indicates an increase in the complexity and diversity of questions, with a notable shift from descriptive to evaluative and exploratory types, as students moved through the stages. Additionally, questions became more focused and multi-variable in nature by the final presentation.

DISCUSSION & CONCLUSION

This study's findings reveal a significant evolution in the data science questions posed by 12th-grade students from populations historically underrepresented in computing across four stages of question collection points. Initially, students focused on simple, descriptive questions, aligning with the early stages of the statistical problem-solving process as defined by the GAISE framework (Arnold & Franklin, 2021). As students gained familiarity with the data and analysis tools, their inquiries became more sophisticated, transitioning towards comparative, evaluative, and exploratory questions. This progression is consistent with research highlighting the importance of scaffolding and student interest in developing inquiry skills (Wiser et al., 2012; Chu et al., 2021).

Importantly, these findings demonstrate the wealth of ideas, interests, and abilities students have for engaging in data science.

The observed evolution also underscores the value of interest-driven, culturally relevant topics in fostering student engagement and facilitating the development of statistical literacy skills (Dolenc & Kazanis, 2020). In our observation, students gained confidence and experience with self-selected topics, particularly those related to their consistent interests in animals and music and, they intended to pose more focused, multi-variable, and evaluative questions, leading to richer insights from their data analysis. This aligns with Dierker et al.'s (2016) findings on the positive impact of fostering student confidence and interest in developing statistical literacy skills crucial for interdisciplinary research.

These insights have significant implications for designing effective data science curricula for students historically underrepresented in computing fields. Educators should provide initial support while progressively introducing more complex analytical tasks, encouraging students to develop their critical thinking and data analysis skills. Tailoring projects to include themes that resonate with students' interests can further enhance their learning experience and engagement in computing education fields.

While this study offers valuable insights, it is limited by the small sample size (15 students in the final presentation submission) and, thus, serves as an initial exploratory investigation of the driving research questions. Future research with larger and more diverse samples could examine the trajectory of question-type changes at the student level and investigate the impact of specific instructional interventions on inquiry skill development.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (Award # 2141655). Any opinions, conclusions, and/or recommendations are those of the investigators and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Arnold, P., & Franklin, C. (2021). What makes a good statistical question?. Journal of Statistics and Data Science Education, 29(1), 122-130.

Arnold, P., & Franklin, C. (2021). What makes a good statistical question?. *Journal of Statistics and Data Science Education*, 29(1), 122-130.

Biehler, R., Veaux, R. D., Engel, J., Kazak, S., & Frischemeier, D. (2022). Research on data science education. *Statistics Education Research Journal*, 21(2), Article 2. https://doi.org/10.52041/serj.v21i2.606

Chu, S. K. W., Reynolds, R. B., Tavares, N. J., Notari, M., & Lee, C. W. Y. (2021). 21st century skills development through inquiry-based learning from theory to practice. Springer International Publishing

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Dasgupta, S. (2016). *Children as data scientists: Explorations in creating, thinking, and learning with data* [Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/107580

Dierker, L., Alexander, J., Cooper, J. L., Selya, A., Rose, J., & Dasgupta, N. (2016). Engaging diverse students in statistical inquiry: a comparison of learning experiences and outcomes of under-represented and non-underrepresented students enrolled in a multidisciplinary project-based statistics course. *International Journal for the Scholarship of Teaching and Learning*, 10(1), n1.

Dolenc, N. R., & Kazanis, W. H. (2020). A potential for interest driven learning to enhance the inquiry based learning process. *Science Educator*, 27(2), 121-128.

Krishnamurthi, S., Schanzer, E., Politz, J. G., Lerner, B. S., Fisler, K., & Dooman, S. (2020). Data science as a route to AI for middle- and high-school students (arXiv:2005.01794). *arXiv preprint arXiv:2005.01794*.

Israel-Fishelson, R., Moon, P., Tabak, R., & Weintrop, D. (2024). Understanding the Data in K-12 Data Science. Harvard Data Science Review, 6(2).

Michaelis, J. E., & Weintrop, D. (2022). Interest development theory in computing education: A framework and toolkit for researchers and designers. ACM Transactions on Computing Education, 22(4), 1-27.

O'Neil, C., & Schutt, R. (2013). Doing data science: Straight talk from the frontline. " O'Reilly Media, Inc.".Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Rivas, C. (2012). Coding qualitative data. In C. Seale (Ed.), *Researching Society and Culture* (pp. 367–392). SAGE Publications. https://eprints.soton.ac.uk/378176/

Saldaña, J. (2016). The coding manual for qualitative researchers. SAGE.

Weiland, T., & Engledowl, C. (2022). Transforming curriculum and building capacity in K–12 data science education. *Harvard Data Science Review*, *4*(4).

Weintrop, D. & Israel-Fishelson, R. (2024). Bringing Students' Lives Into Data Science Classrooms. Harvard Data Science Review, 6(3).

Wiser, M., Smith, C. L., & Doubler, S. (2012). Learning progressions as tools for curriculum development: Lessons from the Inquiry Project. In Learning progressions in science (pp. 357-403). Brill.