DATA SCIENCE EDUCATION K-12

Investigating the Evolution of Interest-Driven Data Science Questions Posed by High-School Students

Xiaoxue Zhou, University of Maryland College Park, xxzhou@umd.edu
Rotem Israel-Fishelson, University of Maryland College Park, rotemisf@umd.edu
David Weintrop, University of Maryland College Park, weintrop@umd.edu
Peter Moon, University of Maryland College Park, pmoon@umd.edu
Yue Xin, University of Maryland College Park, yxin21@umd.edu

Session Description

In today's data-driven world, students must be able to explore and analyze the data surrounding them. A crucial aspect of this process is formulating meaningful research questions that can be addressed with the available data. This study investigates the data science inquiry process of high school students. We analyzed 213 student-generated questions from the final project of an innovative interest-driven data science curriculum. Through a qualitative analytic approach, we examined changes in question types, complexity, and scope across four stages of data collection. The findings shed light on a shift from descriptive to more complex, evaluative, and exploratory questions. It also highlights the importance of providing scaffolding, culturally relevant content, and adaptive instructional strategies in data science education. These elements are essential for empowering students from marginalized backgrounds and fostering their engagement and success in the field.

Background

Data science education aims to equip students with the technical skills to analyze datasets, investigate phenomena, and pursue questions (Weiland & Engledowl, 2022), while also fostering critical thinking, informed decision-making, and advocating for fair data practices (Biehler et al., 2022). It is essential for students to have the ability to formulate meaningful questions that can be explored and answered through data analysis. Recognizing the significance of interests and inquiry, particularly for young learners, as emphasized by the Interest-Driven Computing Education Framework (Michaelis & Weintrop, 2022), this research explores the inquiry process of students from historically excluded populations in computing. Specifically, we investigated the API CAN CODE curriculum, an interest-driven curriculum that introduces students to computing concepts through programming, data analysis, and visualization using public data sources. At the heart of the curriculum is encouragement for students to pose and then attempt to answer questions on topics of their interest. As students' data science skills and knowledge progress, they are invited to revise existing questions and formulate new ones, offering a unique opportunity to observe the evolution of students' data science questions over time. The research question guiding this work is: How do students' data science questions evolve in type, scope, and complexity throughout the project? To answer the question, we collected and analyzed 213 student-generated questions from the four stages to assess changes in question types, complexity, and scope through a qualitative analytic approach.

Methods & Participants

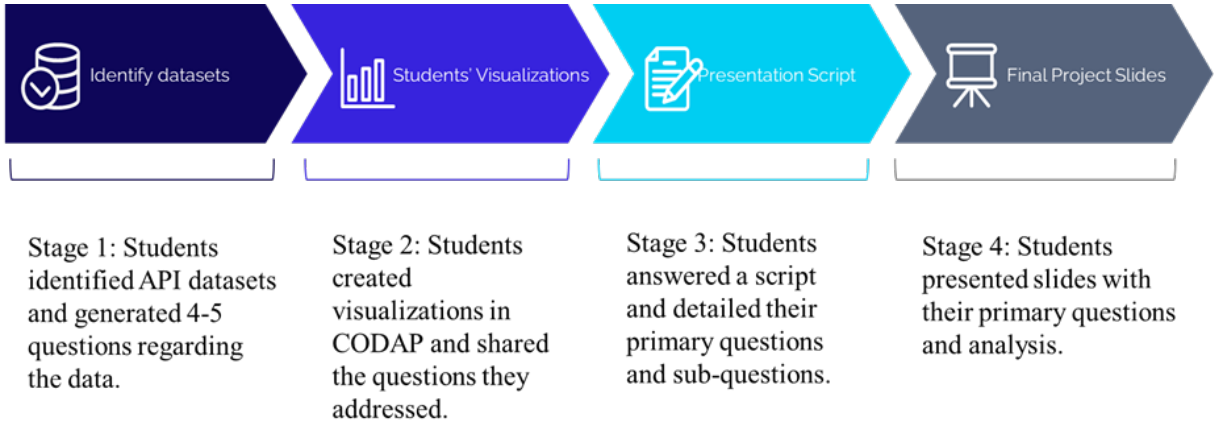
In early 2024, we implemented a three-unit data science curriculum in two 12th-grade classes at a Public Charter High School in the Mid-Atlantic U.S. The curriculum focused on computational foundations of data analysis, and data visualization. In the final project, students chose a topic, crafted questions, identified data sources, and used data science practices to communicate their findings. Twenty-three students consented to participate. 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**:** We collected questions at four stages (Figure 1): dataset identification, visualization, sub-questions for presentation, and final presentation. 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.

Figure 1. Students' Questions Evolution throughout the Four Stages



Data Analysis: We created a codebook and categorized student-generated questions by type, scope, and complexity. Six question types were identified: Descriptive-Attribute, Descriptive-Comparison, Descriptive-Distribution, Exploratory, Predictive, and Evaluative. The scope was categorized as broad or focused, and complexity as single-variable or multi-variable. Table 2 presents these three dimensions, including definitions and examples from the data. Two researchers independently coded the data, 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?"

Results

The students' questions evolved across the four stages of the project (Table 3). At first, students 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). The 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 showed a change 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 increased (from 7.48% to 14.81%), often using terms like "best" or "popular." Student 4's question about Jhené 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, 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 3. 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

Discussion & Conclusion

The study shows a significant progression in students' questions, shifting from simple descriptive inquiries to more complex, evaluative, and exploratory ones. This evolution corresponds with research on the importance of scaffolding and student interest in developing inquiry skills (Wiser et al., 2012; Chu et al., 2021). 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). These insights have implications for educators, suggesting to start by offering initial support and then gradually introduce more complex analytical tasks to encourage the development of students' critical thinking and data analysis skills. 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.

References

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.

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.

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.

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

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.